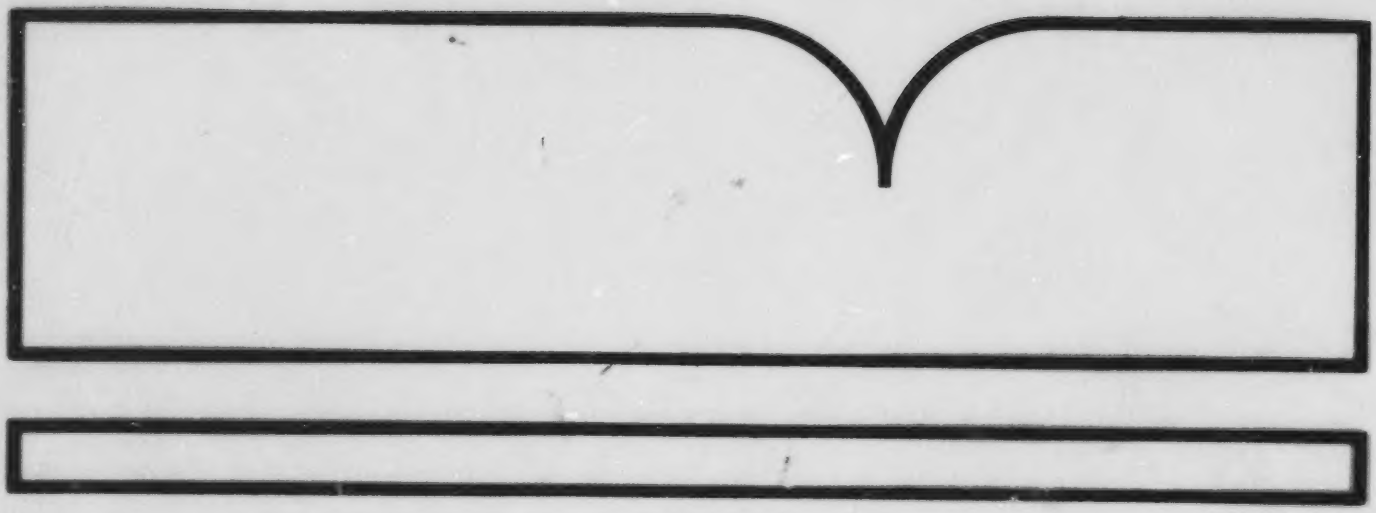
PB85-106060

Interim Procedures for Evaluating Air Quality Models (Revised)

(U.S.) Environmental Protection Agency
Research Triangle Park, NC

Sep 84



U.S. Department of Commerce
National Technical Information Service
NTIS

United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-84-023
September 1984

Air

PB85-106060

EPA

Interim Procedures For Evaluating Air Quality Models (Revised)

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

EPA-450/4-84-023

Interim Procedures for Evaluating Air Quality Models (Revised)

U.S. ENVIRONMENTAL PROTECTION AGENCY
Monitoring and Data Analysis Division
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

September 1984

Disclaimer

This report has been reviewed by The Office of Air Quality Planning and Standards, U. S. Environmental Protection Agency, and has been approved for publication. Mention of trade names or commercial products is not intended to constitute endorsement or recommendation for use.

Preface

The quantitative evaluation and comparison of models for application to specific air pollution problems is a relatively new problem area for the modeling community. Although considerable experience has been gained in applying the procedures contained in an earlier version of this document, it is expected that there will continue to be a number of problems in carrying out the procedures described herein. Thus, procedures discussed in this document should continue to be considered interim.

EPA Regional Offices and State air pollution control agencies are encouraged to use this document to judge the appropriateness of a proposed model for a specific application. However, they must exercise judgment where individual recommendations are not of practical value. After a period of time during which further experience is gained, problem areas will become better defined and will be addressed in additional revisions to this document.

The procedures described herein are specifically tailored to operational evaluation, as opposed to scientific evaluation. The main goal of operational evaluation as applied here is to determine whether a proposed model is that which is most reliable for use in a specific regulatory action. The ability of various sub-models (plume rise, etc.) to accurately reproduce reality or to add basic knowledge assessed by scientific evaluation is not specifically addressed by these procedures.

An example illustrating the procedures described in this document has been prepared, and is attached as Appendix B. As noted in the preface to Appendix B, the primary utility of the example is to illustrate some

considerations in designing the performance evaluation protocol. The example is not intended to be a "model" to be followed in an individual application of these procedures.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA 450/4-84-023		PB85 106060

4. TITLE AND SUBTITLE	5. REPORT DATE
Interim Procedures for Evaluating Air Quality Models (Revised)	September 1984
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Monitoring and Data Analysis Division Office of Air Quality Planning and Standards U.S. Environmental Protection Agency Research Triangle Park, N.C. 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Monitoring and Data Analysis Division Office of Air Quality Planning and Standards U.S. Environmental Protection Agency Research Triangle Park, N.C. 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE EPA-450/4-84-023

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This document describes interim procedures for use in accepting, for a specific regulatory application, a model that is not recommended in the Guideline on Air Quality Models. The procedure involves a technical evaluation and a performance evaluation, utilizing measured ambient data, of the proposed nonguideline model. The primary basis for accepting the proposed model is a demonstration that it performs better (better agreement with measured data) that the guideline model or the model that EPA would normally use in the given situation. The acceptance procedure may also consider the technical merits of the proposed model and, especially in cases where an EPA recommended model cannot be identified, the performance of the model in comparison to a set of specially designed performance standards. A major component of the procedure is the development of a protocol which describes exactly how the performance evaluation will be conducted and what the specific basis for accepting or rejecting the proposed model will be.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Meteorology Mathematical Models Performance Evaluation Performance Standards Statistics	Performance Measures Technical Evaluation	4B 12A

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
Release Unlimited	Unclassified	144
	20. SECURITY CLASS *(This page)*	22. PRICE
	Unclassified	

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

Table of Contents

Preface iii

Table of Contents v

List of Tables vii

List of Figures ix

Summary xi

1.0 INTRODUCTION 1

1.1 Need for Model Evaluation Procedures 2
1.2 Basis for Evaluation of Models 4
1.3 Coordination with Control Agency 5

2.0 PRELIMINARY ANALYSIS 7

2.1 Regulatory Aspects of the Application 7
2.2 Source and Source Environment 8
2.3 Reference Model 10
2.4 Proposed Model 11
2.5 Preliminary Estimates 12
2.6 Technical Comparison with the Reference Model 13
2.7 Technical Evaluation When No Reference Model Is Used 14
2.8 Technical Summary 16

3.0 PROTOCOL FOR PERFORMANCE EVALUATION 19

3.1 Performance Measures 20

3.1.1 Model Bias 21
3.1.2 Model Precision 23
3.1.3 Correlation Analysis 24

3.2 Data Organization 25

3.3 Protocol Requirements 27

3.3.1 Performance Evaluation Objectives 29
3.3.2 Selecting Data Sets and Performance Measures 31
3.3.3 Weighting the Performance Measures 34
3.3.4 Determining Scores for Model Performance 36
3.3.5 Format for the Model Comparison Protocol 38

3.4 Protocol When No Reference Model Is Available 42

4.0 DATA BASES FOR THE PERFORMANCE EVALUATION 45

4.1 On-Site Data 46

4.1.1 Air Quality Data 46
4.1.2 Meteorological and Emissions Data 50

4.2 Tracer Studies 51
4.3 Off-Site Data 53

5.0 MODEL ACCEPTANCE 57

5.1 Execution of the Model Performance Protocol 57
5.2 Overall Acceptability of the Proposed Model 59
5.3 Model Application 60

6.0 REFERENCES 63

APPENDIX A. Reviewer's Checklist A-1

APPENDIX B. Narrative Example B-1

APPENDIX C. Procedure for Calculating Non-Overlapping Confidence Intervals C-1

List of Tables

Number	Title	Page
3.1	Statistical Estimators and Basis for Confidence Limits on Performance Measures	22
3.2	Summary of Candidate Data Sets for Model Evaluation	28
3.3	Summary of Data Sets and Performance Statistics for Various Performance Evaluation Objectives	32
3.4	Suggested Format for the Model Comparison Protocol	39
5.1	Suggested Format for Scoring the Model Comparison	58

List of Figures

Number	Title	Page
1	Decision Flow Diagram for Evaluating a Proposed Air Quality Model	xii
3.1	Observed and Predicted Concentration Pairings Used in Model Performance Evaluations	26

Preceding page blank

Summary

This document describes interim procedures for use in accepting, for a specific application, a model that is not recommended in the Guideline on Air Quality Models[1]. The primary basis for the model evaluation assumes the existence of a reference model which has some pre-existing status and to which the proposed nonguideline model can be compared from a number of perspectives. However for some applications it may not be possible to identify an appropriate reference model, in which case specific requirements for model acceptance must be identified. Figure 1 provides an outline of the procedures described in this document.

After analysis of the intended application, or the problem to be modeled, a decision must be made on the reference model to which the proposed model can be compared. If an appropriate reference model can be identified, then the relative acceptability of the two models is determined as follows. The model is first compared on a technical basis to the reference model to determine if it can be expected to more accurately estimate the true concentrations. Next a protocol for model performance comparison is written and agreed to by the applicant and the appropriate regulatory agency. This protocol describes how an appropriate set of field data will be used to judge the relative performance of the proposed and the reference model. Performance measures recommended by the American Meteorological Society[2] will be used in describing the comparative performance of the two models in an objective scheme. That scheme should consider the relative importance to the problem of various modeling objectives and the degree to which the individual performance measures support those objectives. Once the plan for performance evaluation is

Preceding page blank

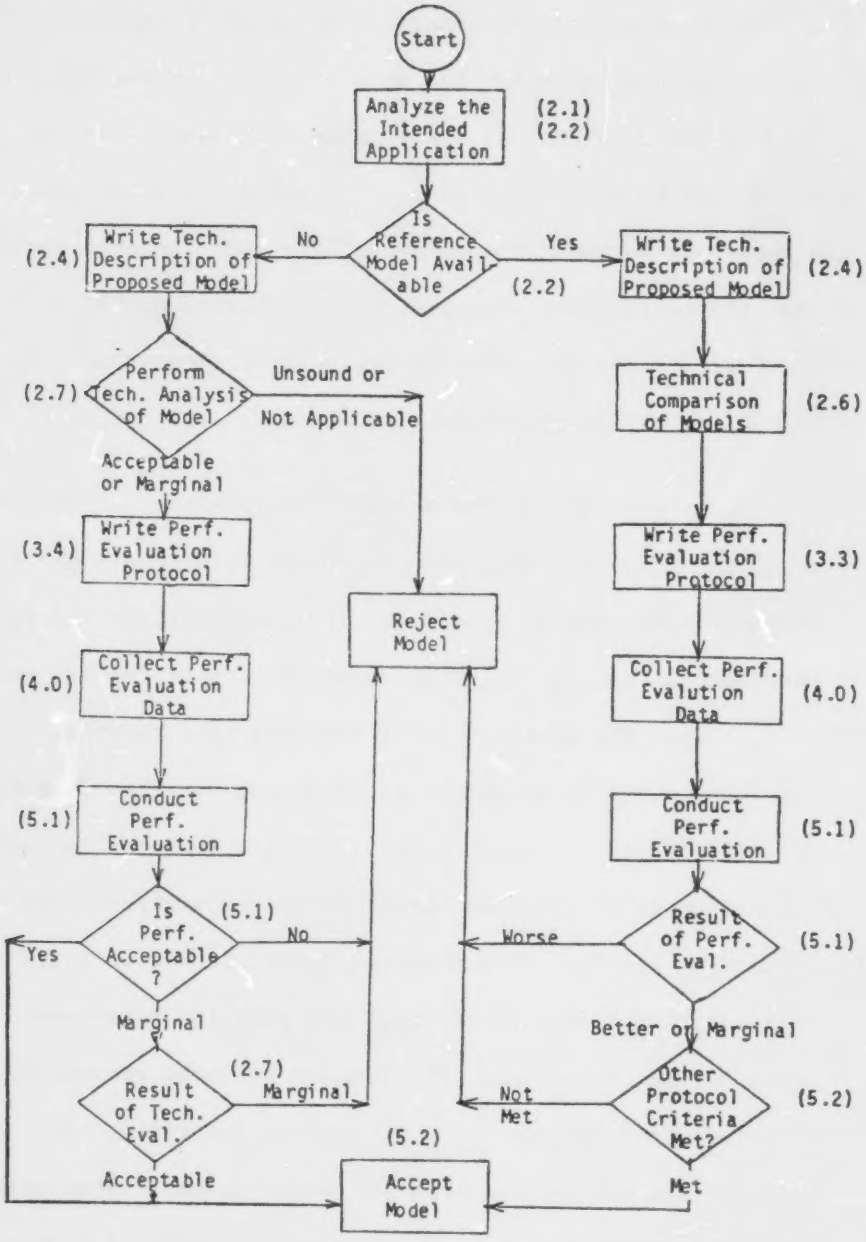


Figure 1. Decision Flow Diagram for Evaluating a Proposed Air Quality Model (Applicable Sections of the Document are indicated in Parentheses.)

written and the data to be used are collected/assembled, the performance measure statistics are calculated and the weighting scheme described in the protocol is executed. Execution of the decision scheme will lead to a determination that the proposed model performs better, worse or about the same as the reference model for the given applications. The final determination on the acceptability of the proposed model should be primarily based on the outcome of the comparative performance evaluation. However, it may also be based, if so specified in the protocol, on results of the technical evaluation, the ability of the proposed model to meet minimum standards of performance, and/or other specified criteria.

If no appropriate reference model is identified, the proposed model is evaluated as follows. First the proposed model is evaluated from a technical standpoint to determine if it is well founded in theory, and is applicable to the situation. This involves a careful analysis of the model features and intended usage in comparison with the source configuration, terrain and other aspects of the intended application. Secondly, if the model is considered applicable to the problem, it is examined to see if the basic formulations and assumptions are sound and appropriate to the problem. (If the model is clearly not applicable or cannot be technically supported, it is recommended that no further evaluation of the model be conducted and that the exercise be terminated.) Next, a performance evaluation protocol is prepared that specifies what data collection and performance criteria will be used in determining whether the model is acceptable or unacceptable. Finally, results from the performance evaluation should be considered together with the results of the technical evaluation to determine acceptability.

INTERIM PROCEDURES FOR EVALUATING AIR QUALITY MODELS

1. INTRODUCTION

This document describes interim procedures that can be used in judging whether a model, not specifically recommended for use in the Guideline on Air Quality Models[1], is acceptable for a given regulatory action. It identifies the documentation, model evaluation and data analyses desirable for establishing the appropriateness of a proposed model.

This document is only intended to assist in determining the acceptability of a proposed model for a specific application (on a case-by-case basis). It is not intended for use in determining which of several models and/or model options (similar to an optimization procedure) are best for application to a given situation, nor does it address procedures to be used in model development and "validation." It is not for use in determining whether a new model could be acceptable for general use and/or should be included in the Guideline on Air Quality Models. This document also does not address criteria for determining the adequacy of alternative data bases to be used in models, except in the case where a nonguideline model requires the use of a unique data base. The criteria or procedures generally applicable to the review of fluid modeling procedures are contained elsewhere.[3,4,5]

The reminder of Section 1 describes the need for a consistent set of evaluation procedures, provides the basis for performing the evaluation, and suggests how the task of model evaluation should be coordinated between the applicant and the control agency. Section 2 describes the preliminary technical analysis needed to define the regulatory problem,

the choice of the reference and proposed models, and the regulatory consequences of applying these models. Section 2 also contains a suggested method of analysis to determine the applicability of the proposed model to the situation. Section 3 discusses the protocol to be used in judging the performance of the proposed model. Section 4 describes the design of the data base for the performance evaluation. Section 5 describes the execution of the performance evaluation and provides guidance for combining these results with other criteria to judge the overall acceptability of the proposed model. Appendix A provides a reviewer's checklist which can be used by the appropriate control agency in determining the acceptability of the applicant's evaluation. Appendix B provides an example illustrating the use of the procedures. Appendix C describes a procedure for calculating non-overlapping confidence intervals.

1.1 Need for Model Evaluation Procedures

The Guideline on Air Quality Models makes specific recommendations concerning air quality models and the data bases to be used with these models. The recommended models should be used in all evaluations relative to State Implementations Plans (SIPs) and Prevention of Significant Deterioration (PSD) unless it is found that the recommended model is inappropriate for a particular application and/or a more appropriate model or analytical procedure is available. However, for some applications the guideline does not recommend specific models and the appropriate model must be chosen on a case-by-case basis. Similarly, the recommended data bases should be used unless such data bases are unavailable or inappropriate. In these cases, the guideline states that other models and/or data bases deemed appropriate by the EPA Regional Administrator may be used.

Models are used to determine the air quality impact of both new and existing sources. The majority of cases where nonguideline models have been proposed in recent years have involved the review of new sources, especially in connection with PSD permit applications. However, most Regional Offices have also received proposals to use nonguideline models for SIP relaxations and for general area-wide control strategies.

Many of the proposals to use nonguideline models involve modeling of point sources in complex terrain and/or a shoreline environment. Other applications have included modeling point sources of photochemical pollutants, modeling in extreme environments (arctic/tropics/deserts), modeling of fugitive emissions and modeling of burning where smoke management (a form of intermittent control) is practiced. For these applications a refined approach is not identified in the Guideline on Air Quality Models. Also a relatively small number of proposals have involved applications where a recommended model is appropriate, but another model is considered preferable.

The types of nonguideline models proposed have included: (1) minor modification of computer codes to allow a different configuration/number of sources and receptors that essentially do not change the estimates from those of the basic model; (2) modifications of basic components in recommended models, e.g., different dispersion coefficients (measured or estimated), wind profiles, averaging times, etc; and (3) completely new models that involve non-Gaussian approaches and/or phenomenological modeling, e.g. temporal/spatial modeling of the wind flow field.

The Guideline on Air Quality Models, while allowing for the use of alternative models in specific situations, does not provide a technical basis for deciding on the acceptability of such techniques. To assure a

more equitable approach in dealing with sources of pollution in all sections of the country it is important that both the regulatory agencies and the entire modeling community strive toward a consistent approach in judging the adequacy of techniques used to estimate concentrations in the ambient air. The Clean Air Act[6] recognizes this goal and states that the "Administrator shall specify with reasonable particularity each air quality model or models to be used under specified sets of conditions . . ."

The use of a consistent set of procedures to determine the acceptability of nonguideline models should also serve to better ensure that the state-of-the-science is reflected. A properly constructed set of evaluation criteria should not only serve to promote consistency, but should better serve to ensure that the best technique is applied. It should be noted that a proposed model cannot be proprietary since it may be subject to public examination and could be the focus of a public hearing or other legal proceeding.

1.2 Basis for Evaluation of Models

The basis for accepting a proposed model for a specific application, as described in this document, involves a comparison of performance between the proposed model and an applicable reference model. The proposed model would be acceptable for regulatory application if its performance is better than that of the reference model. It should not be applied to the problem if its performance were inferior to that of the reference model. This model should also meet other criteria that may be specified in the protocol.

A second basis for accepting or rejecting a proposed model could involve the use of performance criteria written specifically for the intended application. While this procedure is limited by a lack of

experience in writing such criteria and the necessity of considerable subjectivity, it is recognized that in some situations it may not be possible to specify an appropriate reference model. Such a scheme should ensure that the proposed model is technically sound and applicable to the problem. Further, the model should pass certain performance requirements that are acceptable to all parties involved. Marginal performance together with a marginal determination on technical acceptability would suggest that the model should not be used.

At the present time one cannot set down a complete set of objective evaluation criteria and standards for acceptance of models using these concepts. Bases for such objective criteria are lacking in a number of areas, including a consistent set of standards for model performance, scientific consensus on the nature of certain flow phenomena such as interactions with complex terrain, etc. However, this document provides a framework for inclusion of future technical criteria, as well as currently available criteria.

1.3 Coordination with Control Agency

The general philosophy of this document is that the applicant or the developer of the model should perform the analysis. The reviewing agency should review this analysis, perform checks, and/or perform an independent analysis. The reviewing agency must have access to all of the basic information that went into the applicant's analysis (model computer code, all input data, all air quality data) so that an independent judgment is possible.

To avoid costly and time-consuming delays in execution of the model evaluation, the applicant is strongly urged to maintain close

liaison with the reviewing agency(s), both at the beginning and throughout the project. A mininum* of two reports should be submitted to the control agency for review and subsequent negotiation. The first report should contain the preliminary analysis, the protocol for the performance evaluation and the design of the data base network. Before any monitors are deployed or data collection begins, it is important that the control agency concur on all aspects of the planned evaluation, including choice of the reference model, design of the performance evaluation protocol and the design of the data base network. The second report would be submitted at the conclusion of the study. It should describe the data base, the results of executing the protocol, and the model chosen for application.

---

*As a mechanism to maintain close liaison between the source and control agency, the submission of other periodic progress reports is encouraged.

2.0 PRELIMINARY ANALYSIS

As a prerequisite to design of the performance evaluation and the data base network, it is necessary to develop and document a complete understanding of all regulatory and technical aspects of the model application. This preliminary analysis establishes the regulatory requirements of the application and describes the source and its surroundings. Based on these factors, the analysis identifies and describes a reference model or historically based regulatory model which would normally be applied to the source(s). The preliminary analysis includes concentration estimates from the reference and proposed models, based on existing data and appropriate emission rates. If the protocol specifies that the technical analysis is to be considered in the final decision (see Section 3) then the application-specific technical aspects of the two models are compared using techniques described in the Workbook for Comparison of Air Quality Models.[7] This workbook is used to develop a judgment on the scientific credibility of the models for the regulatory application.

The outcome and primary purpose of the preliminary analysis is to provide a focus for the performance evaluation (Section 3) and for identification of the requisite data bases (Section 4). A secondary purpose is to provide a technical basis for judging the model, in the event that the performance evaluation is inconclusive. The preliminary analysis requirements are detailed in the following subsections.

2.1 Regulatory Aspects of the Application

The preliminary analysis should establish the pollutant or pollutants to be modeled, the averaging times, e.g. 3-hour, 24-hour, etc. for these pollutants, and the limiting ambient criteria (standards, PSD

increments, etc.). The current regulatory classification, e.g., attainment, nonattainment, PSD Class I, should be documented. The regulatory boundaries of the area for which concentration estimates apply should also be established. Existing emission limits, if any, should be identified.

For example, there may be a question whether the SO_2 emission limits from several sources in an attainment area can be relaxed, and if so, by how much. In this case, the 3-hour, 24-hour and annual SO_2 ambient air standards apply, as do the Class II increments for these averaging times. There may also be a distant Class I PSD area for which any emission relaxation could result in increment consumption; as such, incremental concentration estimates corresponding to the three averaging times would be required in that area. The allowable time frame for regulatory action should also be identified since the evaluation of model performance involves a significant amount of time and expenditure of resources. Allowable emission rates during the period of model evaluation should be specified.

2.2 Source and Source Environment

To define the important source-receptor relationships involved in a regulatory modeling problem it is necessary to assemble a complete description of the source and its surroundings. Information on the source or sources involved includes the configuration of the sources, location and heights of stacks, stack parameters (flow rates and gas temperature) and location of any fugitive emissions to be included. Existing and proposed emission rates should be identified for each averaging time that corresponds to an ambient air quality standard applicable to pollutants under consideration. In the case of complex industrial sources it is

also generally necessary to obtain a plant layout including dimensions of plant buildings and other nearby buildings/obstacles. Sources should be characterized in as much detail as possible, i.e. commensurate with the input requirements of the models (See Sections 2.3 and 2.4). For example, source emissions should be assembled as mobile and area line source segments, grid squares, etc.

Information on the source surroundings are usually best identified on a topographic map or maps that cover the modeling area. The areal coverage is sometimes predetermined by political jurisdiction boundaries, i.e., an air quality region. More often, however, modeling is confined to the region where any significant threat to the standards or PSD increments is likely to exist. The locations of major existing sources (for the pollutants in question), urban areas, PSD Class I areas, and existing meteorological and air quality data should be identified on the maps.

A determination should be made whether the source(s) in question are located in an urban or rural setting. The recommended procedure for making this determination utilizes the techniques of Auer[8] where the land use within a 3 km radius of the source is classified. Other techniques, based on population and judgmental considerations may be used if they can be shown to be more appropriate.

The method to be used in establishing the ambient concentration due to all other existing sources should be established. If nearby sources are to be modeled, then their emissions and source characterization needs to be specified. Applicable background concentrations and the method used to estimate them should be documented.

2.3 Reference Model

The reference model is the model that would normally be used by the regulatory agency in setting emission limits for the source. The choice of reference model should be made by the appropriate regulatory agency and follow from guidance provided in the Guideline on Air Quality Models.

However, not all modeling situations are covered by recommended models. For example, models for point sources of reactive pollutants or shoreline fumigation problems are not included. In other cases the model normally used by the regulatory agency might be a screening technique that does not lend itself easily to performance evaluations. In these circumstances the applicant and the reviewing agency should attempt to agree on an appropriate and technically defensible reference model, which provides for hour-by-hour estimates based on the current technical literature and on past experience. Major considerations are that the reference model is applicable to the type of problem in question, has been described in published reports or the open literature, and is capable of producing concentration estimates for all averaging times for which a performance measure statistic must be calculated (usually 1-hour and the averaging times associated with the standards/increments). This latter requirement usually* precludes the use of screening techniques which rely on assumed meteorological conditions for a worst case.

Where it is clearly not possible to specify a reference model, the proposed model must "stand alone" in the evaluation. In such cases

---

*Some screening techniques do contain provisions for hour-by-hour estimates and as such they may be used.

the technical justification and the performance evaluation necessary to determine acceptability should be more substantial. Section 2.7 discusses a rationale for determining if the model is technically justified for use in the application. Section 3.4 discusses some considerations in designing the performance evaluation protocol when no reference model comparison is involved.

2.4 Proposed Model

The model proposed for use in the intended application must be capable of estimating concentrations corresponding to the regulatory requirements of the problem as identified in Section 2.1. In order to conduct the performance evaluation, the model should be capable of sequentially estimating hourly concentrations based on meteorological and emission inputs.

A complete technical description of the model is needed for the analysis in Sections 2.6 or 2.7. This technical description should include a discussion of the features of the proposed model, the types of modeling problems for which the model is applicable, the mathematical relationships involved and their basis, and the limitations of the model. The model description should take the form of a report or user's manual that completely describes its operation. Published articles which describe the model are also useful. If the model has been applied to other problems, a review of these applications should be documented. For models designed to handle complex terrain, land/water interfaces and/or other special situations, the technical description should focus on how the model treats these cases. To the maximum extent possible, evidence for the validity of the methodologies should be included.

2.5 Preliminary Estimates

Once the reference and proposed models are identified, it is essential that, at least in a preliminary sense, the consequences of applying each of these models to the regulatory problem be established. The questions to be answered are: (1) What are the preliminary concentration estimates for each model that would be used to establish emission limits? (2) Where are the locations of such critical concentrations? and (3) What are the differences between estimates at these locations? The preliminary estimates should utilize the appropriate emission rates for the regulatory problem and whatever representative meteorological data are available before the evaluation.* In those infrequent cases where no representative meteorological data can be identified, it may be necessary to collect on-site data before making preliminary estimates.

It is recommended that two or three separate preliminary estimates of the concentration field be made. The first set of estimates should be made with the screening techniques mentioned or referenced in the Guideline on Air Quality Models. The second set of estimates should be done with the proposed model and the third set with the reference model. Estimates for all applicable averaging times should be calculated. The three sets of estimates serve to define the modeling domain and critical receptors. They also aid in determining the applicability of the proposed model (Sections 2.6 and 2.7), the development of a performance evaluation protocol, and the design of requisite data networks (Sections 3 and 4).

---

*A final set of model estimates, to be used in decision making, may utilize additional data collected during the performance evaluation as input to the appropriate model.

2.6 Technical Comparison with the Reference Model

When an appropriate reference model can be identified it may prove useful to compare the proposed model with the reference model. Emphasis should be on dispersion conditions and subareas of the modeling domain that are most germane to the regulatory and technical aspects of the problem (Sections 2.1 and 2.2). The procedures described in the Workbook for Comparison of Air Quality Models are appropriate for this comparison. This Workbook contains a procedure whereby a proposed model is qualitatively compared, on technical grounds, to the reference model, and the intended use of the two models and the specific application are taken into account.

The Workbook procedure is application-specific; that is, the results depend upon the specific situation to be modeled. The reference model serves as a standard of comparison against which the user gages the proposed model. The way in which the proposed model treats twelve aspects of atmospheric dispersion, called "application elements," is determined. These application elements represent physical and chemical phenomena that govern atmospheric pollutant concentrations and include such aspects as horizontal and vertical dispersion, emission rate, and chemical reactions. The importance of each element to the application is defined in terms of an "importance rating." Tables giving the importance ratings for each element are provided in the Workbook, although they may be modified under some circumstances. The heart of the procedure involves an element-by-element comparison of the way in which each element is treated by the two models. These individual comparisons, together with the importance ratings for each element in the given application, form the basis upon which the final comparative evaluation of the two models is made.

It is especially important that the user understand the physical phenomena involved, because the comparison of two models with respect to the way that they treat these phenomena is basic to the procedure. Sufficient information is provided in the Workbook to permit these comparisons. Expert advice may be required in some circumstances. If alternate procedures are used to complete the technical comparison of models, they should be discussed with the reviewing agency. The results of the comparison may be used in the overall model evaluation in Section 5.

2.7 Technical Evaluation When No Reference Model Is Used

If it is not possible to identify an appropriate reference model (Section 2.3), then the procedures of Section 2.6 cannot be used and the proposed model must be technically evaluated on its own merits. The technical analysis of the proposed model should attempt to qualitatively answer the following questions:

1. Are the formulations and internal constructs of the model well founded in theory?

2. Does the theory fit the practical aspects and constraints of the problem?

To determine whether or not the underlying assumptions have been correctly and completely stated requires an examination of the basic theory employed by the model. The technical description of the model discussed in Section 2.4 should provide the primary basis for this examination. The examination of the model should be divided into several subparts that address various aspects of the formulation. For example, for some models it might be logical to separately examine the methodologies used to characterize the emissions, the transport, the diffusion, the

plume rise, and the chemistry. For each of these model elements it should be determined whether the formulations are based on sound scientific, engineering and meteorological principles and whether all aspects of each element are considered. Unsound or incomplete specification of assumptions should be flagged for consideration of their importance to the actual modeling problem.

For some models, e.g., those that entail a modification to a model recommended in the Guideline on Air Quality Models or to the reference model, the entire model would not need to be examined for scientific credibility. In such cases only the submodel or modification should be examined. Where the phenomenological formulations are familiar and have been used before, support for their scientific credibility can be cited from the literature.

For models that are relatively new or utilize a novel approach to some of the phenomenological formulations, an in-depth examination of the theory should be undertaken. The scientific support for such models should be established and reviewed by those individuals who have broad expertise in the modeling science and who have some familiarity with the approach and phenomena to be modeled.

To determine how well the model fits the specific application, the assumptions involved in the methodologies proposed to handle each phenomenon should be examined to see if they are reasonable for the given situation. To determine whether the assumptions are germane to the situation, particular attention should be paid to assumptions that are marginally valid from a basic standpoint or those that are implicit and unstated. For assumptions that are not met, it should be established that these deficiencies do not cause significant differences in the

estimated concentrations. The most desirable approach takes the form of sensitivity testing by the applicant where variations made on these assumptions are indeed critical. Such an exercise should be conducted, if possible, and should involve estimates that reflect alternate assumptions before and after modification of formulas or data. However, in many cases this exercise may be too resource consumptive and the proof of model validity should still rest with the performance evaluation described in Section 3.

Execution of the procedures in this section should lead to a judgment on whether the proposed model is applicable to the problem and can be scientifically supported. If these criteria are met, the model can be designated as appropriate and should be applied if its field performance (Section 5) is acceptable. When a model cannot be supported for use based on this technical evaluation, it should be rejected. When it is found that the model could be appropriate, but there are questionable assumptions, then the model may be designated as marginal and carried forward through the performance evaluation.

2.8 Technical Summary

The final step in the technical analysis is to combine the results of Sections 2.1 through 2.6/2.7 into a technical summary. This summary should serve to define (1) the scope of the issues to be resolved by the performance evaluation, (2) the areal and temporal extent of the differences in estimates between the proposed and the reference models, and (3) the reasons why the two models produce different estimates and/or different concentration patterns.

The technical summary provides a focus for the performance evaluation and the design of the requisite data base network. The results of the technical summary are used in Section 3 to establish criteria for the performance evaluation protocol and in Section 4 to define the requisite data base.

3.0 PROTOCOL FOR PERFORMANCE EVALUATION

The goal of the model performance evaluation is to determine whether the proposed model provides better estimates of concentrations germaine to the regulatory aspects of the problem than does the reference model. To achieve this goal, model concentration estimates are compared with observed concentrations in a variety of ways.* The primary methods of comparison produce statistical information and constitute a statistical performance evaluation.

This section describes a procedure for evaluating the performance of the proposed model and for determining whether that performance is adequate for the specific application. The procedure requires that a protocol be prepared for comparing the performance of the reference and proposed models. The protocol must be agreed upon by the applicant and appropriate regulatory agencies prior to collection of the requisite data bases. The description of the protocol includes a scheme to (1) weight the relative importance of various performance measures to the regulatory goals of the evaluation and (2) objectively discriminate between the relative performance of the proposed and reference models. Some guidance is also provided on how to write a protocol and evaluate model performance when comparison with a reference model is not possible.

Before going into the details of the protocol, it is important to review briefly some of the statistical performance measures that are commonly used to assess the performance of a model against measured data. It is also useful to consider how the ambient data base is commonly broken down into

---

*Concentration and meteorological data needed for the performance evaluation are discussed in Section 4. The data base network design/requirements are partially determined by the nature of and amount of performance statistics defined in the protocol.

Preceding page blank

data subsets which are operated on by the statistical measures. Section 3.1 describes these performance measures and Section 3.2 briefly describes some of the commonly used data subsets.

Model performance should be evaluated for each of the averaging times specified in the appropriate regulation(s). In addition, performance for models whose basic averaging time is shorter than the regulatory averaging time should also be evaluated for that shorter period, provided, of course, that the measurements are available for shorter averaging periods. For example, a model may calculate sequential 1-hour concentrations for SO_2 from which concentrations for longer averaging periods can be computed. Performance of this model can thus be evaluated separately for 1-, 3-, and 24-hour averages and, if appropriate, for the annual mean. It should be noted that although frequency distribution statistics are indicated in Table 3.1, they may be considered somewhat redundant when performance measures of both bias and precision are used. For this reason, graphical displays of the cumulative frequency distribution of observed and predicted values may be useful as supplementary aids in the evaluation process.

3.1 Performance Measures

The basic tools used in determining how well a model performs in any given situation are performance measures. Performance measures can be thought of as surrogate quantities whose values serve to characterize the discrepancy between predictions and observations. Values obtained from applying the performance measures to a given data base are most often statistical in nature; however, certain performance measures (e.g., frequency distributions) may be more qualitative than quantitative in nature.

Performance measures may be classified as magnitude of difference measures and correlation or association measures. Magnitude of difference measures present a quantitative estimate of the discrepancy between measured concentrations and concentrations estimated by a model at the monitoring sites. Correlation measures quantitatively delineate the degree of association between estimations and observations.

Table 3.1 lists a number of the more commonly used and recommended performance statistics for model evaluation purposes. These statistics and the corresponding nomenclature are taken from Fox[2] and are based primarily on the recommendations of an AMS Workshop on Dispersion Model Performance held in 1980. Since the statistics and basis for confidence limits are described extensively in most statistical texts, only a brief description of how these measures apply to model performance is presented below. Although each of the statistics provide a quantitative measure of model performance, they are somewhat easier to interpret when accompanied by graphical techniques such as histograms, isopleth analyses and scatter diagrams.

3.1.1 Model Bias

Many of the performance statistics serve to characterize, in a variety of ways, the behavior of the model residual, defined as the observed concentration minus the estimated concentration. For example, model bias is determined by the value of the model residual averaged over an appropriate range of values. Large over- and underestimations may cancel in computing this average. Supplementary information concerning the distribution of residuals should therefore be supplied. This supplementary information consists of confidence intervals about the mean value, and histograms or frequency distributions of model residuals.

Table 3.1 Statistical Estimators and Basis for Confidence Limits on Performance Measures

Performance Measure To Be Estimated	Paired Concentrations		Unpaired Concentrations	
	Estimator	Basis for Confidence Interval	Estimator	Basis for Confidence Interval
Bias	$\bar{d}$	One sample "t" with adjustments for autocorrelation	$\bar{C}_o - \bar{C}_p$	Two sample "t"
	$d_{0.5}$	Wilcoxon matched pair	$\bar{C}_o - \bar{C}_p$	Mann-Whitney
Noise	s^2_d	λ^2	$s^2_{C_o} / s^2_{C_p}$	F - statistic
Gross Variability	$\frac{\Sigma d^2}{N}$	N/A (1)		Note: No separate Noise and Gross Variability are estimated
Absolute Deviations	$\lvert\bar{d}\rvert$	N/A (1)	$\lvert\bar{C}_o - \bar{C}_p\rvert$	N/A (1)
Correlations	r	Fisher "z" transform	N/A (2)	N/A (2)
Frequency Distributions	F(d)	K-S Statistic Test for Normality	$F(C_o) - F(C_p)$	K-S Statistic

(1) Not Applicable - Statistic can be calculated but not meaningful for this type of analysis.

(2) Not Applicable - No statistic exists for this type of data.

For certain applications, especially cases in which the proposed model is designed to simulate concentrations occurring during important meteorological processes, it is important to estimate model bias under different meteorological conditions. The degree of data disaggregation is a compromise between the desired goals of defining a large enough number of meteorological categories to cover a wide range of conditions and having a sufficient number of observations in each category to calculate statistically meaningful values. For example, it may be appropriate to stratify data by lumped stability classes, unstable (A-C), neutral (D) and stable (E-F) rather than by individual classes A, B, C, D, E and F. The use of wind speed classes may also be appropriate.

3.1.2 Model Precision

Model precision refers to the average amount by which estimated and observed concentrations differ as measured by a different type of residual than that used for bias, that is the absence of an algebraic sign. While large positive and negative residuals can cancel when model bias is calculated, the unsigned residuals comprising the precision measures do not cancel. Thus, they provide an estimate of the error scatter about some reference point. This reference point can be the mean error or zero error. Two types of precision measures are the noise, which delineates the error scatter about the mean error, and the gross variability, which delineates the error scatter about zero error.

The performance measure for noise is either the variance of the residuals, s_d^2, or the standard deviation of the residuals, s_d. The performance measure for gross variability is the mean square error, or the root-mean-square-error. An alternate performance measure for the gross variability is the mean absolute residual, $\overline{|d|}$. The mean

absolute residual is statistically more robust than the root-mean-square-error; that is, it is less affected by removal of a few extreme values.

Supplementary analyses for model precision should include confidence limits, as appropriate, and computation of these measures for selected meteorological categories as discussed in Section 3.1.1.

3.1.3 Correlation Analyses

Correlation analyses involve parameters calculated from linear least squares regression and associated graphical analyses. The numerical results constitute quantitative measures of the association between estimated and observed concentrations. The graphical analyses constitute supplementary qualitative measures of the same information. There are three types of correlation analyses; coupled space-time, spatial, and temporal analyses.

Coupled space-time correlation analysis involves computing the Pearson's correlation coefficient, r, or an equivalent nonparametric coefficient such as Spearman's ρ or Kendall's τ. The parameters a and b of the linear least squares regression equation should be included. A scattergram of the predicted data pairs is supplementary information which should be presented.

Spatial correlation analysis involves calculating the spatial correlation coefficient and presenting isopleth analyses of the predicted and observed concentrations for particular periods of interest. The spatial coefficient measures the degree of spatial alignment between the estimated and observed concentrations. The method of calculation involves computing the correlation coefficient for each time period and determining an average over all time periods. Estimates of the spatial

correlation coefficient for single source models are most reliable for calculations based on data intensive networks such as those contained in a tracer study. Isopleths of the distributions of estimated and observed concentrations for periods of interest should be presented and discussed.

Temporal correlation analysis involves calculating the temporal correlation coefficient and presenting time series of observed and estimated concentrations or of the model residual for each monitoring location. The temporal correlation coefficient measures the degree of temporal alignment between observed (C_o) and predicted (C_p) concentrations. The method of calculation is similar to that for the spatial correlation coefficient. Time series of C_o and C_p or of model residuals should be presented and discussed for each monitoring location.

3.2 Data Organization

The performance measures described above may be applied to various combinations of observed and predicted values depending on the objectives of the evaluation and the nature of the regulatory problem (i. e., the intended application). For example, when "once per year" ambient standards are of primary concern, observed and predicted maximum (or near maximum) concentrations should be compared. Since complete space and/or time pairing is often not important from a regulatory point of view, the appropriate data combination need not be restricted to only concentration pairs having the same hour or location.

There are many possible combinations of observed and predicted concentrations that may be chosen for evaluation. Thus, it is useful to organize, at least conceptually, the complete data set into a matrix of observed and predicted values as exhibited in Figure 3.1. Entries in the center of the figure are completely paired in time and space. Entries

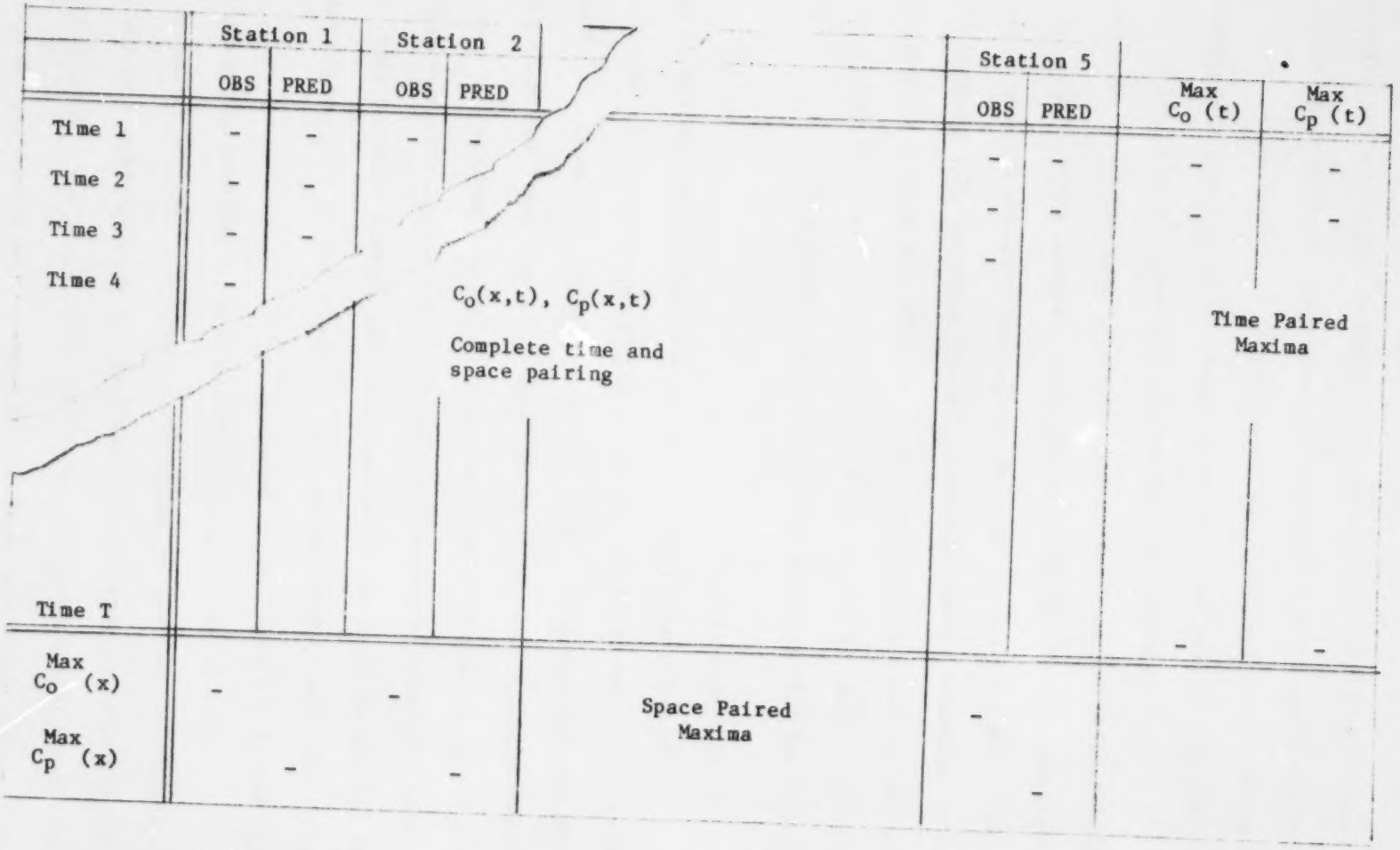


Figure 3.1 Observed and Predicted Concentration Pairings Used in Model Performance Evaluations

shown in the bottom two rows and last two columns represent, are respectively, pairs of maximum concentration paired in space only and time only.

Since the figure permits illustration of only a few data combinations which may be of interest from a regulatory viewpoint, a more complete tabulation of data combinations (data sets) is shown in Table 3.2. The first type of data combination refers to "peak concentration" which by definition excludes the low concentration comparisons. Except for combination A-3 (completely paired peak residuals), all data sets involve some degree of spatial and/or temporal unpairing between observed and predicted values. The second type of data combination refers to "all concentrations" which comprise complete time and space pairing for all predicted and observed values within a defined category. For example, data set B-1 refers to the set of all observed and predicted values at a given station, paired in time. Since each station is evaluated separately, the total number of data combinations is equal to the total number of stations.

The rationale for selecting particular data combinations and statistics to evaluate various aspects of model performance is simplified by first establishing major objectives to be accomplished by the performance evaluation. The procedure for establishing these objectives and for assigning levels of importance to each objective is discussed in the following section.

3.3 Protocol Requirements

Because of the variety of statistical measures and data combinations that might be considered for evaluation purposes, it is essential that a written protocol be prepared and agreed to by the applicant and appropriate control agency before the data collection and evaluation

Table 3.2. Summary of Candidate Data Sets for Model Evaluation

A. Peak Concentration Comparisons	B. All-Concentrations Comparisons
(A-1) Compare highest observed value for each event with highest prediction for same event (paired in time, not location).	(B-1) Compare observed and predicted values at a given station, paired in time. (A data set for each station.)
(A-2) Compare highest observed value for the year at each monitoring station with the highest prediction for the year at the same station (paired in location, not time).	(B-2) Compare observed and predicted values for a given time period, paired in space (not appropriate for data sets with few monitoring sites).
(A-3a) Compare maximum observed value for the year with highest predicted values representing different time or space pairing (fully unpaired, paired in location, paired in time, paired in space and time).	(B-3) Compare observed and predicted values at all stations, paired in time and location (one data set) and by time of day.
(A-3b) Compare maximum predicted value for the year with highest observed values for various pairings, as in (A-3a).	(B-4) Same as (B-3), but for subsets of events by meteorological conditions (stability and wind speed) and by time of day.
(A-4a) Compare highest N (=25) observed and highest N predicted values, regardless of time or location.	
(A-4b) Compare highest N (=25) observed and highest N predicted values, regardless of time, for a given monitoring location. (A data set for each station.)	
(A-5) Same as (A-4a), but for subsets of events by meteorological conditions (stability and wind speed) and by time of day.	

process is initiated. Conceptually, the protocol describes how various performance measures will be used to compare the relative performance of the proposed and reference models in a manner that is most relevant to the regulatory need (the intended application as described in Section 2.1.) To organize this concept it is suggested that the protocol contain four major components as follows: (1) a definition of the performance evaluation objectives to be accomplished in terms of their relevance to the regulatory application; (2) a compilation of specific data sets and performance measures that will be applied under each performance objective; (3) an objective scheme for assigning weights to each performance measure and data set combination; and (4) an objective scheme for scoring the performance of the proposed model relative to the reference model.

This section discusses the factors to be considered in establishing such a protocol for an individual performance evaluation. Although some experience has been gained in applying the techniques to actual regulatory situations, it remains clear that the procedures described below must remain general enough to adequately cover all types of regulatory problems.

3.3.1 Performance Evaluation Objectives

The first step in developing the model performance protocol is to translate the regulatory purposes associated with the intended modeling application into performance evaluation objectives, which, in turn, can be linked to specific performance measures and data sets. This step is important since each intended modeling application is unique with respect to source configuration, the critical source-receptor relationships, the types of ambient levels to be protected (e.g., NAAQS vs. PSD), averaging

times of concern (e.g., 1-hr, 3-hr, 24-hr, etc.), and the form of the ambient standard (e.g., not to be exceeded more than once per year vs. annual).

In most applications, the primary regulatory purposes can be stated clearly in terms that relate directly to certain performance measures and data sets. For example, if the primary regulatory purpose is to prevent violations of short-term ambient standards which might be threatened by construction of a large isolated SO_2 source, then the ability of the models to accurately predict highest 3-hour and 24-hour concentrations is critical. In this example situation, the primary performance objective might be stated as: "Determine the accuracy of peak estimates in the vicinity of the proposed plant."

While "peak accuracy" is the first order objective in this example, other performance objectives can be stated that relate to the ability of the selected models to perform over a variety of concentrations levels and conditions. For example, additional confidence can be placed in a model if it is also accurate in estimating the magnitude of lower concentrations at specific stations and for specific meteorological events. Thus, a second order objective might be stated as "Determine the accuracy of estimates of concentrations over a range of concentrations, time periods, and stations."

A third performance evaluation objective which can be derived from this example regulatory application is related to measures of spatial and temporal correlation. While a model may adequately predict peaks and average levels at given stations, a measure of additional confidence can be gained if the model also traces the time sequence of

concentrations reasonably well. Thus, a third order performance evaluation objective might be stated as "Determine the degree of correlation between predicted and measured values in time and space." While correlation is a reasonably stringent performance measure (time and/or space pairing is required), it is ranked below the previous two performance objectives. Even good correlation can be obtained in cases where the magnitude of peak levels are poorly predicted and for which a large overall bias exists.

It should be noted that the generic formulation and number of performance objectives for any given application may differ substantially from those illustrated here. In other words, the specific regulatory purpose should be the guide for the selection of those performance objectives that are most directly relevant to the intended application.

3.3.2 Selecting Data Sets and Performance Measures

Once the performance evaluation objectives are established, it is necessary to choose among the various data combinations and performance statistics listed in Tables 3.1 and 3.2. These are used to characterize the ability of the models to meet the evaluation objective. Table 3.3 summarizes the more important data sets and performance statistics relevant to each generic objective described above. These objectives have been arbitrarily numbered in relative order of importance as they might pertain to the hypothetical SO_2 regulatory application described above. For an actual application, any of the three generic objectives (or some other derived objective) could have a higher level of importance depending on the nature of the regulatory problem.

Table 3.3 shows, for each performance evaluation objective, a suggested list of the most relevant data sets and performance measures

Table 3.3. Summary of Data Sets and Performance Statistics for Various Performance Evaluation Objectives

Performance Evaluation Objectives	Data Sets (Table 3.2)	Performance Statistics (Table 3.1)	Supplementary Graphics
1. Determine Model Accuracy for Peak Values	A-3a, A-3b	Single Valued Residuals	None
	A-4a, A-4b, A-5	$s^2_{C_o}/s^2_{C_p}$, $\bar{d}$	Freq Dist of Top 25
	A-1, A-2	s^2_d, $\bar{d}$	Freq Dist of All Values
2. Determine Model Accuracy Over Entire Concentration Domain	B-1,B-2,B-3,B-4	s^2_d, $\bar{d}$	Selected Isopleths and Time Series Plots
3. Spatial and Temporal Correlation	B-1,B-2,B-3,B-4	r, Regression Statistics	Scattergrams
	A-1,A-2	r, Regression Statistics	Scattergrams

Notes: (1) If particular site(s) are crucial (i.e., PSD), then analyses should be confined to a site or a subset of important sites.

(2) For reactive pollutants, performance measures should be developed for each of a number of selected days.

along with supplementary graphical displays that may prove useful in the evaluation process. For example, the first order objective shown as "Accuracy of Peak Values" has three data sets. Except for selected cases, these data sets correspond to the peak concentration category shown in Table 3.2. While each data set offers some measure of information regarding accuracy of peak estimates, the focus is on different aspects of peak levels that may be of greater importance in some applications. Data sets A-3a and A-3b relate most directly to short-term ambient standards. However, they suffer by being statistically non-robust compared to data sets A-4 and A-5 which involve a greater number of highest values in the computation of the performance measures. Data set A-1, since it consists of a large number of values (one pair for each time period), is subject to the least statistical variability but suffers by including many events that may be below the concentration range of primary concern. Thus, the tradeoff is between the degree of confidence desired and the degree of regulatory relevance associated with each candidate data set/performance statistic.

The performance statistics are directly tied to the nature of the performance evaluation objective and the degree of natural pairing between the measured and predicted values. Since the first and second objectives both relate to accuracy, measures of bias and precision (noise and variability) are indicated. The third performance evaluation objective, by virtue of its definition, involves correlation and hence the correlation coefficient, r, is indicated. Note that whenever performance measures are applied to paired data (e.g., A-1, A-2), the measure of precison is the noise, s_d, while for unpaired data (e.g., A-4a, A-4b, A-5), the ratio of variances is indicated.

A precise procedure for choosing data sets and performance statistics for each objective cannot be illustrated here. It must be determined by consideration of the nature of the regulatory purpose(s), the degree of confidence desired in the final result and the resources available for the evaluation. In specific applications, some of the statistics and/or data sets may be omitted depending upon the degree of redundancy or relevance to the regulatory problem. For example, data set B-3 uses all available pairs of data but requires that only one set of statistics be calculated. This contrasts with data set B-1 which also makes use of all data pairs but requires a separate calculation for each station. The decision regarding the use of both data sets (i.e., B-3 and B-1) depends to some extent on the need to know how well the models perform at specific station locations over all concentration levels.

3.3.3 Weighting the Performance Measures

Once the appropriate performance evaluation objectives, data sets and performance measures are specified, it is necessary to establish the relative importance each performance measure should hold in the final decision scheme. It is suggested that the relative importance of the performance measures be objectively established by assigning weights to the performance evaluation objectives and also to each performance measure according to how well that measure characterizes the objective. The assignment of weights in any given situation is somewhat judgmental and may differ slightly among trained analysts. Thus, it is important in the protocol to document the rationale used to establish the relative weights. It is suggested that, in order to keep the problem simple, that weights be established on the basis of a percentage or fraction of a total 100 points.

Generally the first order objective would be weighted most heavily while less important objectives would be weighted less heavily depending upon their importance in the application. As an example, the first order objectives might be weighted 50 percent, the second order objectives 30 percent, and the third order objectives 20 percent. For each performance objective, each combination of performance measures and data sets must also be given a weight. Again the determination of the appropriate weight for each performance measure is judgmental and should be accompanied by a rationale. Some of the judgments involved, for example are: (1) Is model bias a more important factor than gross variability? (2) Is accurate prediction of the magnitude of the peak more important than accurate prediction of the location of that peak? Answers to these questions vary with the application and will result in different assignment of weights accordingly. Those measures of performance which best characterize the ability of either model to more accurately estimate the concentrations that are critical to decision-making should carry the most weight. If the estimated maximum concentration controls the emission limit for the source(s), then more weight should be given to performance measures that assess the models' ability to accurately estimate the maxium concentration.

The magnitude of the weights should also take into consideration the degree of confidence that can reasonably be assigned to the performance statistic to be calculated, i.e., only minimal confidence can be placed in single-valued residuals since these values are non-robust and sensitive to unusual conditions. Generally, there will be some trade-off between degree of confidence and relevance of the particular performance measure. This means that the most relevant performance

measures may be given less weight than otherwise might be assigned were confidence in the result not a critical factor.

3.3.4 Determining Scores for Model Performance

The final step in writing the protocol is to establish how each performance statistic (calculated by applying a performance measure to a given data set) can be translated into a performance evaluation score. Such a scheme involves definition of the rationale to be used in determining the degree to which each pair of performance measure statistics supports the advantage of one model over the other. Stated differently, it is necessary to have a measure of the degree to which better performance of one model over the other can be established for each performance measure. It seems apparent that the more confidence one has that one model is performing better than the other, the higher that model should score in the final decision on the appropriateness of using that model. Clearly this is important when at least one of the models is performing moderately well. For example, if only one model appears to be unbiased, the degree to which the other is biased can be a factor in quantifying the relative advantage of the apparently unbiased model.

Qualitatively, the problem of determining which model is performing better is straightforward. Clearly, the model with the smaller residuals, the smaller bias, the smaller noise and the higher correlation coefficient is better. The difficulty, which is not straightforward, is how to meaningfully quantify the comparative advantage that one model has over the other. There are several approaches that can be used.

In one approach, a "score" is derived for each pair (one for each model) of performance statistics. The number of points which are awarded is based on the degree of statistical significance attached

to the difference in each model's ability to reproduce the observed data. The level of significance could be determined by the degree to which confidence limits on performance measures of each model overlap or, alternatively, on an hypothesis-testing method in which a specified confidence level is assigned. A procedure for awarding points using confidence limits is outlined in the Appendix B. In the "example problem" positive points are awarded for each performance statistic if the proposed model performs better than the reference model and negative points if the reference model performs better. The (absolute) magnitude of the score is dependent on the relative difference in the model's performance of each model but is limited to the "maximum score" established for each measure. Such a maximum score is directly proportional (or perhaps equivalent) to the weight for each measure.

The reader is cautioned that the actual level of statistical significance is based to a varying degree on the assumptions that model residuals are independent of one another, an assumption that is clearly not true. For example, model residuals from adjacent time periods (e.g., hour-to-hour) are known to be positively correlated. Also, the proposed and reference model residuals for a given time period are related since the residual for each model involves the same observed concentration for a given data pairing. However, if such statistical limitations are recognized, this approach can be useful as a quantitative indicator for determining which model is performing better in a particular situation.

A second approach for assigning points is to assign points separately to each model for each performance measure; then, by difference, derive a net point total for the proposed model; the point

total can be positive or negative, as discussed below. Various schemes, both statistical and nonstatistical, have been proposed for assigning points based on the numerical difference between measured and predicted levels (i.e., the performance measures). A predetermined function of the performance measure could be used to award points for each model. The number of possible points could range from zero when the model performs unacceptably (e.g., the bias exceeds the observed average by more than 50 percent) up to a maximum when the model performs perfectly (e.g., the bias is zero). The net number of points assigned to the proposed model would then be the number of points awarded to the proposed model minus the number of points awarded to the reference model. A positive difference favors the proposed model while a negative difference favors the reference model. In essence, this second approach involves a subjective decision as to what constitutes acceptable performance. Although this suggests a "de facto" performance standard, the result may be informative since the total accumulated points for each model would serve as an indicator of how poorly or how well the models are doing overall in terms of the particular application.

3.3.5 Format for the Model Comparison Protocol

A suggested format for the model comparison protocol, based on the weighting and scoring scheme discussed above, is provided by Table 3.4. The example format in Table 3.4 is for the first order performance objectives. A similar format would be used for second order objectives, third order objectives, etc. An example of "filled out" tables using this format are provided in Tables 1-4 of Appendix B.

In the first column of Table 3.4 the various data sets or subsets which will be used to generate statistical or other information

Table 3.4. Suggested Format for the Model Comparison Protocol

Part A. First Order Objective: _____

Data Set (Sub Set)	Pairing Space	Pairing Time	Performance Measures	Method for Awarding Points*	Averaging Times	Maximum Points	Rationale
1._____	a._____	_____	a._____	_____	_____	_____	_____
					_____	_____	
					_____	_____	
	b._____	_____	b._____	_____	_____	_____	_____
					_____	_____	
					_____	_____	
	c.		c.				
	'		'				
	'		'				
2._____							

*E.g., confidence levels, ratios, etc.

are listed. The second column specifies the various combinations of time and space pairing between estimates and measurements. The third column lists the performance measures to be employed on each data set and time/space pairing. The fourth column contains the numerical scheme that will be used to determine the points to be awarded to the proposed model. The fifth column lists the averaging times for which statistical or other information that will be obtained and the sixth column lists the maximum points or "weighting" for each statistic (or other objective quantity). In the last column a rationale is to be provided for the choices made in the preceding columns.

This format is intended to provide a quick visual summary of the overall scheme for scoring the relative performance of the models and for use in establishing criteria for selecting the best model. The actual scoring should proceed in a straightforward manner once the performance statistics have been calculated and used to allocate points for each indicated data set. A total score can be derived by simply summing the individual scores which will result in a net positive score if the proposed model scores higher and a net negative score if the reference model scores higher.

Although it is tempting to choose the higher scoring model for use in the regulatory application, two additional criteria may be considered in arriving at the final determination. First, it may be desirable to establish, a priori, standards of performance that must be met before either model may be selected. For example, a limit (positive or negative) on peak bias could be set that, if exceeded, would be sufficient for rejecting the proposed model.

A second selection criteria that may be considered is establishment of a scoring point range that serves to separate outcomes that clearly favor the proposed or reference model and outcomes that do not clearly favor either model. When the score falls within the scoring ranges or "window" where neither model is clearly favored, the final rejection or acceptance of the proposed model could be decided by the outcome of the technical evaluation (Section 2.6 or 2.7). Under this scheme a marginal outcome of the performance evaluation coupled with a marginal or unfavorable outcome of the technical evaluation would suggest that the model not be accepted. Conversely, if the proposed model is clearly technically well founded or superior to the reference model but its performance score falls in the window it probably should be accepted. Several factors might influence the width of such a scoring margin including the representativeness and completeness of the data base and the need to choose a model having a clear performance edge.

If any or all of the above suggested additional criteria are to be considered, then these criteria and their objective use in the decision process need to be specified in the protocol. This requirement is in concert with the basic philosophy of this document that the entire decision-making scheme is specified "up-front" before any data are collected/analyzed which might provide insight into the possible outcome.

After the model selection process is completed it is still desirable to ensure that the chosen model will not underpredict measured concentrations to the extent that the emission limit inferred from application of the model would likely result in violations of the NAAQS or PSD increments. This could occur in those cases where one model outscores the other, and thus judged to be the better performer, yet it still underpredicts

the highest concentrations. To cover such an eventuality it may be desirable to include criteria in the protocol that allow the emission limits or the model to be adjusted to such an extent that attainment of the ambient criteria will be ensured.

3.4 Protocol When No Reference Model Is Available

When no reference model is available, it is necessary to write a different type of protocol based on case-specific criteria for the model performance. However, at the present time, there is a lack of scientific understanding and consensus of experts necessary to provide a supportable basis for establishing such criteria for models. Thus the guidance provided in this subsection is quite general in nature. It is based primarily on the presumption that the applicant and the regulatory agency can agree to certain performance attributes which, if met, would indicate within an acceptable level of uncertainty that the model predictions could be used in decision-making.

A set of procedures should be established based on objective criteria that, when executed, will result in a decision on the acceptability of the model from a performance standpoint. As was the case for the model comparison protocol, it is suggested that the relative importance of the various performance measures be established. Table 3.3 may serve as a guide. However, the performance score for each measure should be based on statistics of d, or the deviation of the model estimates from the true concentration, as indicated by the measured concentrations. For each performance measure, criteria should be written in terms of a quantitative statement. For example, it might be stated that the average model bias should not be greater than plus or minus X at the Y percent significance level. Some considerations in writing such criteria are:

(1) Conservatism. This involves the introduction of a purposeful bias that is protective of the ambient standards or increments, i.e., overprediction may be more desirable than underprediction.

(2) Risk. It might be useful to establish maximum or average deviation from the measured concentrations that could be allowed.

(3) Experience in the performance of models. Several references in the literature[9,10,11,12] describe the performance of various models. These references can serve as a guide in determining the performance that can be expected from the proposed model, given that an analogy with the proposed model and application can be drawn.

As was the case for the model comparison protocol, a decision format or table analogous to Table 3.4 should be established. Execution of the procedures in the table may lead to a conclusion that the performance is acceptable, unacceptable or marginal.

4.0 DATA BASES FOR THE PERFORMANCE EVALUATION

This section describes interim procedures for choosing, collecting and analyzing field data to be used in the performance evaluation. In general there must be sufficient accurate field data available to adequately judge the performance of the model in estimating all the concentrations of interest for the given application.

Three types of data can be used to evaluate the performance of a proposed model. The preferred approach is to utilize meteorological and air quality data from a specially designed network of monitors and instruments in the vicinity of the sources(s) to be modeled (on-site data). In some cases especially for new sources, it is advantageous to use on-site tracer data from a specifically designed experiment to augment or be used in lieu of long-term continuous data. Occasionally, where an appropriate analogy to the modeling problem can be identified, it may be possible to utilize off-site data to evaluate the performance of the model.

As a general reference for this section the criteria and requirements contained in the Ambient Monitoring Guidelines for Prevention of Significant Deterioration (PSD)[13] should be used. Much of the information contained in the PSD monitoring guideline deals with acquiring information on ambient conditions in the vicinity of a proposed source, but such data may not entirely fulfill the input needs for model evaluation.

All data used as input to the air quality model and its evaluation should meet standard requirements or commonly accepted criteria for quality assurance. New site-specific data should be subjected to a quality assurance program. Quality assurance requirements for criteria pollutant measurements are given in Section 4 of the PSD monitoring guideline. Section 7 of the PSD monitoring guideline describes quality

Preceding page blank

assurance requirements for meteorological data. For any time periods involving missing data, it should be specified how such time periods, e.g. data substitution, will be handled.

4.1 On-Site Data

The preferable approach of performance evaluation is to collect an on-site data base consisting of concurrent measurements of emissions, stack gas parameters, meteorological data and air quality data. Given an adequate sample of these data, an on-site data base designed to evaluate the proposed model relevant to its intended application should lead to a definitive conclusion on its applicability. The most important goal of the data collection network is to ensure adequate spatial and temporal coverage of model input and air quality data.

4.1.1 Air Quality Data

The analysis performed in Section 2 serves to define the requisite areal and temporal coverage of the data base and the range of meteorological conditions over which data must be acquired. Once the scope of the data base is established the remaining problem is to define the density of the monitoring network, the specific locations of ambient monitors and the period of time for which data are to be recorded. In general it can be said that the type and quantity of data to be collected must be sufficient to meet the needs of the protocol developed from the guidance provided in Section 3. This determination is a judgment that must be made in advance of the network design; some more specific considerations are now provided.

The number of monitors needed to adequately conduct a performance evaluation is often the subject of considerable controversy. It has been argued that one monitor located at the point of maximum concentration for each averaging time corresponding to the standards or increments should be sufficient. However, the points of maximum concentration are not known but are estimated using the models that are themselves the subject of the performance evaluation, which of course unacceptably compromises the evaluation. It is possible that the use of data from one or two monitors in a performance evaluation may actually be worse than no evaluation at all since no meaningful statistics can be generated. Attempts to rationalize this problem may lead to erroneous conclusions on the suitability of the models. When the data field is sparse, confidence bands on the residuals for the two models will be broad. As a consequence, the probability of statistically distinguishing the difference between the performance of the two models may be unacceptably low.

At the other extreme is a large number of monitors, perhaps 40 or more. The monitors may cover the entire modeling domain or area where significant concentrations, above a small cutoff, can be reasonably expected. The monitors may be sufficiently dense that the entire concentration field (isopleths) is established. Such a concentration field allows the calculation of the needed performance statistics and given adequate temporal coverage, as discussed below, would likely result in narrow confidence bands on the model residuals. With these narrow confidence bands it is easier to distinguish between the relative capabilities of the proposed model vs. the reference model. However, costs associated with such a network would likely be large.

Thus, the number of monitors needed to conduct a significantly meaningful performance evaluation should be judged in advance. Some other factors that should be considered are:

1. Models or submodels that are designed to handle special phenomena should only be evaluated over the spatial domain where that phenomena would result in significant concentrations. Thus, the monitoring network should be concentrated in that area, perhaps with a few outlying monitors for a safety factor.

2. In areas where the concentration gradient is expected to be high (based on preliminary estimates) a high density of monitors should be considered, while in areas of low concentration gradient a less dense network is often adequate.

3. If historical on-site air quality and/or meteorological data are available, these data should also be used to define the locations and coverage of monitors.

In the temporal sense some of the above rationale are also appropriate. A short-term study may lead to low or no confidence on the ability of the models (proposed and reference) to reproduce reality. A multi-year effort will yield several samples and model estimates of the second-highest short-term concentrations, thus providing some basis for a statistically significant comparison of models for this frequently critical estimate. Realistically, multi-year efforts usually have prohibitive costs and one has to rely on somewhat circumstantial evidence, e.g. the upper end of the frequency distribution, to establish confidence in the models' capabilities to reproduce the second-highest concentration.

In general, the data collected should cover a period of record that is truly representative of the site in question, taking into account

variations in meteorological conditions, variations in emissions and expected frequency of phenomena leading to high concentrations. One year of data is normally the minimum, although short-term studies are sometimes acceptable if the results are representative and the appropriate critical concentrations can be determined from the data base. Thus short-term studies are adequate if it can be shown that "worst case conditions" are limited to a specific period of the year and that the study covers that period. Examples might be ozone problems (summer months), shoreline fumigation (summer months) and certain episode phenomena.

Models designed to handle special phenomena need only have enough temporal coverage to provide an adequate (produce significant statistical results) sample of those phenomena. For example, a downwash algorithm might be evaluated on the basis of 50 or so observations in the critical wind speed range.

It is important that the data used in model development or model selection be independent of those data used in the performance evaluation. In most cases, this is not a problem because the model is either based on general scientific principles or is based on air quality data from an analogous situation. However, in some semi-empirical approaches where site-specific levels of pollutants are either an integral part of the model or are used to select certain model options, an independent set of data must be used for performance evaluation. Such an independent data set may be collected at the same site as the one used in model development, but the data set should be separated in time, e.g. use one year of data for model development/tuning and a second year for performance evaluation purposes.

For air quality measurements used in the performance evaluation, it is necessary to distinguish between (1) the contribution of sources that are included in the model and (2) the contribution attributable to background (or baseline levels). The Guideline on Air Quality Models discusses some methods for estimating background. Considerable care should be taken in estimating background so as not to bias the performance evaluation.

4.1.2 Meteorological and Emissions Data

Requisite supporting data such as meteorological and emissions data should be collected concurrently with the ambient data. The degree of temporal resolution of such data should be comparable to that of the ambient data (usually 1-hour) or shorter if model input needs so require. The location and type of meteorological sensors are generally defined by the model input requirements. The more accurately one can pinpoint the location of the plume(s) the less noise that will occur in the model residuals. This can be done by increasing the spatial density and degree of sophistication in meteorological input data, for models that are capable of accepting such data. Continuous collection of representative meteorological input data is important. If multiple (redundant) sensors are to be deployed, a statement should be included in the protocol as to how these data will be used in the performance evaluation.

Accurate data on emissions and stack gas parameters, over the period of record, diminishes the noise in the temporal statistics. The more accurate the emissions data are, the less noise in the residuals. Although data contained in a standard emissions inventory can sometimes be used, it is generally necessary to obtain and explicitly model with

real time (concurrent with the air quality data used in performance evaluation) emissions data. "In-stack" monitoring is highly recommended to ensure the use of emission rates and stack gas parameter data comparable in time to measured ground-level concentrations.

4.2 Tracer Studies

The use of on-site tracer material to simulate transport and dispersion in the vicinity of a point or line source has received increasing attention in recent years as a methodology for evaluating the performance of air quality simulation models. This technique is attractive from a number of standpoints:

1. It allows the impacts from an individual source to be isolated from those of other nearby sources which may be emitting the same pollutants;

2. It is generally possible to have a reasonably dense network of receptors in areas not easily accessible for placement of a permanent monitor;

3. It allows a precise definition of the emission rate;

4. It allows for the emissions from a proposed source to be simulated.

There are some serious difficulties in using tracers to demonstrate the validity of a proposed model application. The execution of the field study is quite resource intensive, especially in terms of manpower. Samplers need to be manually placed and retrieved after each test and the samples need to be analyzed in a laboratory. Careful attention must be placed on quality control of data and documentation of meteorological conditions. As a result most tracer studies are conducted as a short-term

(a few days to a few weeks) intensive campaign where large amounts of data are collected. If conducted carefully, such studies provide a considerable amount of useful data for evaluating the performance of the model. However, the performance evaluation is limited to those meteorological conditions that occur during the campaign. Thus, while a tracer study may allow for excellent spatial coverage of pollutant concentrations, it provides a limited sample, biased in the temporal sense, and leaves an unanswered question as to the validity of the model for all points on the annual frequency distribution of pollutants at each receptor.

Another problem with tracer studies is that the plume rise phenomena may not be properly simulated unless the tracer material can be injected into the gas stream from an existing stack. Thus, for new sources where the material is released from some kind of platform, the effects of any plume rise submodel cannot be evaluated.

Given these problems, the following criteria should be considered in determining the acceptability of tracer tests:

1. The tracer samples should be easily related to the averaging time of the standards in question;

2. The tracer data should be representative of "worst case meteorological conditions";

3. The number and location of the samplers should be sufficient to ensure measurement of maximum concentrations;

4. Tracer releases should represent plume rise under varying meteorological conditions;

5. Quality assurance procedures should be in accordance with those specified or referenced in the PSD monitoring guideline, as well as other commonly accepted procedures for tracer data;

6. The on-site meteorological data base should be adequate;

7. All sampling and meteorological instruments should be properly maintained;

8. Provisions should be made for analyzing tracer samples at remote locations and for maintaining continuous operations during adverse weather conditions, where necessary.

Of these criteria, items 1 and 2 are the most difficult to satisfy because the cost of the study precludes collection of data over an annual period. Because of this problem it is generally necessary to augment the tracer study by collecting data from strategically placed monitors that are operated over a full year. The data are used to establish the validity of the model in estimating the second-highest short term and the annual mean concentration. Although it is preferable to collect these data "on-site," this is usually not possible where a new plant is proposed. It may be possible to use data collected at a similar site, in a model evaluation as discussed in the next subsection.

As with performance evaluations using routine air quality data, sufficient meteorological data must be collected during the tracer study to characterize transport and dispersion input requirements of the model. Since tracer study data are difficult to interpret, it is suggested that the data and methodologies used to collect the data be reviewed by individuals who have experience with such studies.

4.3 Off-Site Data

Infrequently, data collected in another location may be sufficiently representative of a new site so that additional meteorological and air quality data need not be collected. The acceptability of such data rests

on a demonstration of the similarity of the two sites. The existing monitoring network should meet minimum requirements for a network required at the new site. The source parameters at the two sites should be similar. The source variables that should be considered are stack height, stack gas characteristics and the correlation between load and climatological conditions.

A comparison should be made of the terrain surrounding each source. The following factors should be considered:

1. The two sites should fall into the same generic category of terrain:

    a. flat terrain;

    b. shoreline conditions;

    c. complex terrain;

        (1) three-dimensional terrain elements, e.g., isolated hill,
        (2) simple valley,
        (3) two dimentional terrain elements, e.g., ridge, and
        (4) complex valley.

2. In complex terrain the following factors should be considered in determining the similarity of the two sites:

    a. aspect ratio of terrain, i.e., ratio of
        (1) height of valley walls to width of valley,
        (2) height of ridge to length of ridge, and
        (3) height of isolated hill to width of hill base;

    b. slope of terrain;

    c. ratio of terrain height to stack/plume height;

    d. distance of source from terrain, i.e., how close to valley wall, ridge, isolated hill;

    e. correlation of terrain feature with prevailing winds;

    f. the relative size (length, height, depth) of the terrain features.

It is very difficult to secure data sets with the above emission configuration/terrain similarities. Nevertheless, such similarities are of considerable importance in establishing confidence in the representativeness of the performance statistics. The degree to which the sites and emission configuration are dissimilar is a measure of the degree to which the performance evaluation is compromised.

More confidence can be placed in a performance evaluation which uses data collected off-site if such data are augmented by an on-site tracer study (See Section 4.2). In this case the considerations for terrain similarities still hold, but more weight is given to the comparability of the two sets of the observed concentrations. On-site tracer data can be used to test the ability of the model to spatially define the concentration patterns if a variety of meteorological conditions are observed during the tracer tests. Off-site data must be adequate to test the validity of the model in estimating maximum concentrations.

5.0 MODEL ACCEPTANCE

This section describes interim criteria which can be used to judge the acceptability of the proposed model for the specific regulatory application. This involves execution of the performance protocol which will lead to a determination as to whether the proposed model performs better than the reference model. Or when no reference model is available the proposed model may be found to perform acceptably, marginally, or unacceptably in relation to established site-specific criteria. Depending on the results of the performance evaluation, the overall decision on the acceptability of the model might also consider the results of the technical evaluation of Section 2.

5.1 Execution of the Model Performance Protocol

Execution of the model performance protocol involves: (1) collecting the performance data to be used (Section 4); (2) calculation and/or analysis of the model performance measures (Section 3.1); and (3) combining the results in the objective manner described in the protocol (Section 3.3 or Section 3.4) to arrive at a decision on the relative performance of the two models.

Table 5.1 provides a format which may be used to accommodate the results of the model comparison protocol described in Section 3.3.5. If a different protocol format is prepared, it should have the same goal, i.e., to arrive at a decision on how the proposed model performs relative to the reference model.

The first column lists the performance objectives. The next three columns in Table 5.1 are analogous to the first three columns in Table 3.4. The fifth column contains the actual score for each modeling

Preceding page blank

Table 5.1 Suggested Format for Scoring the Model Comparison

	Data Set (Sub Set)	Pairing Space-Time	Performance Measures	Averaging Times	Score	Statistics, Analysis and Calculations that Support the Score
A. First Order Objective:	1.	a.	a.			
		b.	b.			
		c.	c.			
	2.					
B. Second Order Objective						
C. Third Order Objective						
Total Score						

objective as well as the sub-scores for each supporting performance measure. The scores in this column cannot exceed the maximum scores allowed in the protocol. The last column is for the statistics, graphs, analyses and calculations that determine the score for each performance measure, although most of this information would probably be in the form of attachments.

5.2 Overall Acceptability of the Proposed Model

Until more objective techniques are available, it is suggested that the final decision on the acceptability of the proposed model be based primarily on the results of the performance evaluation. The rationale is that the overall state of modeling science has many uncertainties regardless of what model is used, and that the most weight should be given to actual proven performance. Thus when a proposed model is found to perform better than the reference model, it should be accepted for use in the regulatory application. If the model performance is clearly worse than that of the reference model, it should not be used. Similarly, if the performance evaluation is not based on comparison with a reference model, acceptable performance should imply that the model be accepted, while unacceptable performance would indicate that it is inappropriate.

As mentioned at the end of Section 3.3.5, the protocol may contain other criteria, beyond the simple consideration of the score, to determine whether a proposed model is acceptable. For example, the protocol might specify that when the results of the performance evaluation are marginal or inconclusive, the results of the technical evaluation discussed in Section 2 should be used as an aid to deciding on the overall acceptability. In this case, a favorable (better than the reference model) technical review would suggest that the model be used, while a marginal or worse

determination would indicate that the model offers no improvement over existing reference techniques. If Section 2.7 were used to determine technical acceptability, a marginal or inconclusive determination on scientific supportability combined with a marginal performance evaluation would suggest that the model not be applied to the regulatory problem.

Also, as mentioned in Section 3.3.5 the protocol might also specify standards of performance or provisions to guard against underprediction of critical concentrations. If so, these additional criteria must be compared against the performance of the model (in the manner specified in the protocol) before a final decision on the acceptability of the model can be made.

5.3 Model Application

If, as a result of execution of the procedures described in this document, the proposed model is found to be acceptable, then the model should be appropriately applied to the intended application. The data base requirements, the requirements for concentration estimates and other applicable regulatory constraints described in the Guideline on Air Quality Models should be considered.

Much of the data collected during the performance evaluation may also be used during the application phase. For example the meteorological data records may be used as model input. However, in order to ensure that temporal variation of critical meteorological conditions are adequately accounted for, it may be necessary to include a longer period of record. Source characterization data collected during the performance evaluation can be used to the extent that they reflect operating conditions corresponding to the proposed emission limits.

The "proven" model is only applicable for the source-receptor relationship for which the performance evaluation was carried out. Any new application, even for a similar source-receptor relationship, in a different location would generally require a new evaluation. Significant differences in the source configuration, e.g., doubling the stack height from those in existence during the model technical test, may necessitate a new evaluation.

6.0 REFERENCES

1. Environmental Protection Agency. "Guideline on Air Quality Models," EPA-450/2-78-027, Office of Air Quality Planning and Standards, Research Triangle Park, N.C., April 1978.

2. Fox, D. G. "Judging Air Quality Model Performance," Bull. Am. Meteor. Soc. 62, 599-609, May 1981.

3. Environmental Protection Agency. "Guideline for Use of Fluid Modeling to Determine Good Engineering Practice Stack Height," EPA 450/4-81-003, Office of Air Quality Planning and Standards, Research Triangle Park, N.C., July 1981.

4. Environmental Protection Agency. "Guideline for Fluid Modeling of Atmospheric Diffusion," EPA 600/8-81-008, Environmental Sciences Research Laboratory, Research Triangle Park, N.C., April 1981.

5. Environmental Protection Agency. "Guideline for Determination of Good Engineering Practice Stack Height (Technical Support Document for Stack Height Regulations)," EPA 450/4-80-023, Office of Air Quality Planning and Standards, Research Triangle Park, N.C., July 1981.

6. U. S. Congress. "Clean Air Act Amendments of 1977," Public Law 95-95, Government Printing Office, Washington, D.C., August 1977.

7. Environmental Protection Agency. "Workbook for Comparison of Air Quality Models," EPA 450/2-78-028a, EPA 450/2-78-028b, Office of Air Quality Planning and Standards, Research Triangle Park, N.C., May 1978.

8. Auer, A. H., "Correlation of Land Use and Cover with Meteorological Anomalies," J. Appl. Meteor. 17, 636-643, May 1978.

9. Bowne, N. E. "Preliminary Results from the EPRI Plume Model Validation Project--Plains Site." EPRI EA-1788-SY, Project 1616, Summary Report, TRC Environmental Consultants Inc., Wethersfield, Connecticut, April 1981.

10. Lee, R. F., et. al. "Validation of a Single Source Dispersion Model," Proceedings of the Sixth International Technical Meeting on Air Pollution Modeling and Its Application, NATO/CCMS, September 1975.

11. Mills, M. T., et. al. "Evaluation of Point Source Dispersion Models," EPA 450/4-81-032, Teknekron Research, Inc. September 1981.

12. Londegran, R. J., et. al. "Study Performed for the American Petroleum Institute--An Evaluation of Short-Term Air Quality Models Using Tracer Study Data," Submitted by TRC Environmental Consultants, Inc. to API, October 1980.

13. Environmental Protection Agency. "Ambient Monitoring Guideline for Prevention of Significant Deterioration (PSD)," EPA 450/4-80-012, Office of Air Quality Planning and Standards, Research Triangle Park, N.C., November 1980.

Preceding page blank

Appendix A

Reviewer's Checklist

Preface

Each proposal to apply a nonguideline model to a specific situation needs to be reviewed by the appropriate control agency which has jurisdiction in the matter. The reviewing agency must make a judgment on whether the proposed model is appropriate to use and should justify this judgment with a critique of the applicant's analysis or with an independent analysis. This critique or analysis normally becomes part of the record in the case. It should be made available to the public hearing process used to justify SIP revisions or used in support of other proceedings.

The following checklist serves as a guide for writing this critique or analysis. It essentially follows the rationale in this document and is designed to ensure that all of the required elements in the analysis are addressed. Although it is not necessary that the review follow the format of the checklist, it is important that each item be addressed and that the basis or rationale for the determination on each item is indicated.

Preceding page blank

CHECKLIST FOR REVIEW OF MODEL EVALUATIONS

I. Technical Evaluation

A. Is all of the information necessary to understand the intended application available?

1. Complete listing of sources to be modeled including source parameters and location?

2. Maps showing the physiography of the surrounding area?

3. Preliminary meteorological and climatological data?

4. Preliminary estimates of air quality sufficient to (a) determine the areas of probable maximum concentrations, (b) identify the probable issues regarding the proposed model's estimates of ambient concentrations and, (c) form a partial basis for design of the performance evaluation data base?

B. Is the reference model appropriate?

C. Is enough information available on the proposed model to understand its structure and assumptions?

D. Was a technical comparison of the proposed and reference models conducted?

1. Were procedures contained in the Workbook for Comparison of Air Quality Models followed? Are deviations from these procedures supportable or desirable?

2. Are the comparisons for each application element complete and supportable?

3. Do the results of the comparison for each application element support the overall determination of better, same or worse?

Preceding page blank

E. For cases where a reference model is not used, is the proposed model shown to be applicable and scientifically supportable?

II. Model Performance Protocol

A. Are all the performance measures recommended in the document to be used? For those performance measures that are not to be used, are valid reasons provided?

B. Is the relative importance of performance measures stated?

1. Have performance evaluation objectives that best characterize the regulatory problem been properly chosen and objectively ranked?

2. Are the performance measures that characterize each objective appropriate? Is the relative weighting among the performance measures supportable?

C. How are the performance measure statistics for the proposed and the reference model to be compared?

1. Are significance criteria used to discriminate between the performance of the two models established for each performance measure?

2. Is the rationale to be used in scoring the significance criteria supportable?

3. Is the proposed "scoreboard" associated with marginal model performance supported?

4. Are there appropriate performance limits or absolute criteria which must be met before the model could be accepted?

D. How is performance to be judged when no reference model is used?

1. Has an objective performance protocol been written?

2. Does this protocol establish appropriate site-specific performance criteria and objective techniques for determining model performance relative to these criteria?

3. Are the performance criteria in keeping with experience, with the expectations of the model and with the acceptable levels of uncertainty for application of the model?

III. Data Bases

A. Are monitors located in areas of expected maximum concentration and other critical receptor sites?

B. Is there a long enough period of record in the field data to judge the performance of the model under transport/dispersion conditions associated with the maximum or critical concentrations?

C. Are the field data completely independent of the model development data?

D. Where off-site data are used, is the situation sufficiently analogous to the application to justify the use of the data in the model performance evaluation?

E. Will enough data be available to allow calculation of the various performance measures defined in the protocol? Will sufficient data be available to reasonably expect that the performance of the model relative to the reference model or to site-specific criteria can be established?

IV. Is the Model Acceptable

A. Was execution of the performance protocol carried out as planned?

B. Is the model acceptable considering the results of the performance evaluation and the technical evaluation?

Appendix B

Narrative Example

Preface

This narrative example was developed to illustrate the use of the Interim Procedures for Evaluating Air Quality Models. Although the example substantially abbreviates many of the tasks involved in a real model comparison problem and recommended in the interim procedures, it does illustrate the task with which users are most unfamiliar, i.e., the development and execution of the performance evaluation protocol. The following comments/caveats are in order to help better understand and utilize the example:

1. The preliminary technical/regulatory analysis of the intended model application, while included in the example, is significantly foreshortened from that which would normally be needed for an actual problem.

2. The example was specifically designed to illustrate in a very general way the components of the decision making process and the protocol for performance evaluation. As such, the protocol incorporates a broad spectrum of performance statistics with associated weights. The number of statistics contained in this example is probably overly broad for most performance evaluations and perhaps, even for the problem illustrated. Thus its use is not intended to be a "model" for actual situations encountered. For an individual performance evaluation it is recommended that a subset of statistics be used, tailored to the performance evaluation objectives of the problem. The statistical performance measures and associated weighting scheme should be kept as simple (and as understandable) as possible. Complexity implies more precision than exists in the performance measures

Preceding page blank

and weighting schemes and does not reflect the current level of knowledge and experience in conducting performance evaluations.

3. Similarly, the method used to assign scores to each performance statistic (non-over-lapping confidence intervals) is not intended to be a "model" to be followed but should be viewed as only one of several possible techniques to accomplish the same goal.

4. The example does not illustrate the design of the field measurement program required to obtain model evaluation data.

The original narrative example was developed in 1982 by TRC Inc., under contract to EPA. This revised example was adapted from the TRC contract report to reflect the revisions made in the Interim Procedures in September 1984.

Table of Contents

Preface B-3

Table of Contents B-5

List of Tables B-7

List of Figures B-9

1.0 Introduction B-11

2.0 Preliminary Analysis B-13

3.0 Model Evaluation Protocol B-19

3.1 NAAQS Attainment B-19

3.2 PSD Analysis B-31

4.0 Field Measurements B-35

5.0 Performance Evaluation Results and Model Selection B-37

5.1 Results for Model Performance Comparison in the NAAQS Analysis B-46

5.2 Results for Model Performance Comparison in the PSD Analysis B-47

6.0 Summary B-49

7.0 References B-51

List of Tables

Number	Title	Page
1	Model Comparison Protocol for NAAQS Analysis. First-Order Objective: Predicted Highest Concentrations	B-21
2	Model Comparison Protocol for NAAQS Analysis. Second-Order Objective: Predict the Domain of Concentrations	B-24
3	Model Comparison Protocol for NAAQS Analysis. Third-Order Objective: Predict the Pattern (Spatial and Temporal) of Concentrations	B-26
4	Model Comparison Protocol for PSD Analysis. First-Order Objective: Predict Highest Concentrations in PSD Area	B-32
5	Model Comparison Results for NAAQS Analysis. First-Order Objective: Predict Highest Concentrations	B-38
6	Model Comparison Results for NAAQS Analysis. Second-Order Objective: Predict the Domain of Concentrations	B-41
7	Model Comparison Results for NAAQS Analysis. Third-Order Objective: Predict the Pattern (Spatial and Temporal) of Concentrations	B-44
8	Model Comparison Results for PSD Analysis. First-Order Objectives: Predict Highest Concentrations in PSD Areas	B-45

Preceding page blank

List of Figures

Number		Page
1	Field Monitoring Network Near the Clifty Creek Power Plant	B-15
2	Example of Overlapping 95% Confidence Intervals on Bias for Two Models	B-30
3	Example of "Tightened" Confidence Intervals to Result in Non-Overlapping Biases.....	B-30

Preceding page blank

1.0 Introduction

The Interim Procedures for Evaluating Air Quality Models,[1] provide a methodology to judge whether a proposed model, not specifically recommended for use by the Guideline on Air Quality Models,[2] is acceptable for a particular regulatory application. This example model evaluation illustrates the methodology set forth in the Interim Procedures.

The Interim Procedures provide a basis for objectively selecting between the proposed model and a reference model that is either recommended in the Guideline on Air Quality Models or is otherwise agreed to be acceptable for the particular application. To judge which model is more acceptable, the technical features of the two models are compared and then a site-specific performance evaluation of both models is carried out. (For certain regulatory applications, EPA does not designate a reference model. In these cases the Interim Procedures provide a method for assessing the suitability of the proposed model, based on a technical review of the model's applicability and a model performance evaluation).

This example application illustrates the use of the Interim Procedures to select between a proposed model and the reference model for a specific regulatory application. The proposed model is AQ40, a hypothetical air quality dispersion model defined for the purpose of this narrative example. The reference model will be selected as a step in applying the Interim Procedures. The regulatory issue of interest is the short-term air quality impact from a coal-fired power plant in relation to maintaining the National Ambient Air Quality Standards (NAAQS) and the Prevention of Significant Deterioration (PSD) requirements. The Interim Procedures specify the following major steps:

Preceding page blank

1. Perform a preliminary technical/regulatory analysis of intended model application. This includes a definition of the regulatory issues of concern, a description of the source and physical situation being modeled, identification of the appropriate reference model, identification and technical description of the proposed model, preliminary estimates of air quality impacts of the two models and an application specific technical comparison of the proposed and reference models.

2. Prepare a model performance evaluation protocol which specifies the statistical performance comparisons for selecting the appropriate model.

3. Describe the proposed field measurements program required to obtain model evaluation data.

4. Carry out the field measurements program, conduct the performance evaluation of the proposed and reference models with the data collected in the field measurements program using the statistical performance measures specified in the protocol and, based upon an objective comparison of performance results, select/reject the proposed model.

Although each of these steps is discussed in sequence in the narrative example, resources precluded a rigorous illustration of Steps 1 and 3. Thus the primary utility of the example is the detailed illustration of Steps 2 and 4, the design and execution of the performance evaluation.

2.0 Preliminary Analysis

The first step in applying the Interim Procedures for Evaluating Air Quality Models is to analyze the regulatory issues, physical setting and pollutant source to which the proposed and reference models are to be applied. The regulatory requirements dictate the impact region and the averaging periods of interest for the model applications. The physical setting and source characteristics are the basis for selecting the appropriate reference model for the comparative model evaluations. Additionally, preliminary modeling estimates of the expected air quality impacts are made at this time. These estimates are used subsequently in designing the performance evaluation data network and the statistical performance evaluation field program methodology.

The regulatory issue addressed in this example evaluation is the short-term (3- and 24-hour average) air quality impact from a coal-fired power plant located in the Midwest. The power plant used for this example is the Clifty Creek generating station located in southern Indiana and operated by the Indiana-Kentucky Electric Corporation. Compliance with NAAQS and PSD Class I requirements for 3- and 24-hour average sulfur dioxide (SO_2) impacts is the specific regulatory concern to be addressed. No actual PSD Class I region exists in the vicinity of the Clifty Creek station; therefore, a hypothetical Class I region 15 kilometers northeast of the plant is assumed for this example. To assess compliance with NAAQS. model prediction of the highest, second-highest concentration per year within 50 kilometers of the Clifty Creek station is required. PSD regulations are based on the predicted highest, second-highest impact per year of the source within the Class I region.

The physical setting for this example case is the region surrounding the Clifty Creek generating station. The plant is located in the Ohio River Valley in southern Indiana. The Clifty Creek station is a baseload facility and has three 208-meter stacks, with combined average emissions of 8600 g/s SO_2. The average exit temperature is approximately 445°K and the exit velocity ranges from approximately 25 to 50 m/s, depending on the load. The plant is on a flood plain located on a bend in the river. Bluffs rise approximately 60 meters along the Ohio River near the plant. The terrain beyond the bluffs from south-southwest clockwise to north-northeast is quite flat. In the other directions are several streams cutting down to the Ohio River which have created dendritic drainage valleys. The maximum relief in the area (plant grade to the highest monitor, located on a ridge) is about 130 meters and is associated with a ridge between stream cuts. The terrain surrounding Clifty Creek is not "ideally flat"; however, terrain is well below stack height. The site-specific monitoring program includes a 60-meter meteorological tower to measure winds and vertical temperature gradients, and six SO_2 stations 3 to 16 kilometers from the stack. A map of the monitoring network is included as Figure 1.

Selection of the reference model for this application is based upon the recommendations of the Guideline on Air Quality Models. The Guideline recommends the CRSTER model as appropriate for point sources with collocated stacks located in regions where the terrain does not exceed stack height. For this example evaluation the CRSTER-equivalent model, MPTER, cited in the Guideline on Air Quality Models, will serve as the reference model. (Unlike CRSTER, MPTER permits the user to specify receptor locations exactly.)

Monitors	Plant Distance (km)	Elevation MSL(m)	Azimuth From North of Plant(°)
1. Bacon Ridge	15.0	277	40
2. Rykers Ridge	7.4	274	56
3. North Madison	4.5	267	16
4. Hebron Church	11.6	273	24
5. Liberty Ridge	3.1	253	174
6. Canip Creek	8.0	146	99
Clifty Creek Plant Elevation		143	
Elevation at Top of Stacks		351	

Figure 1. Field monitoring network near the Clifty Creek Power Plant.

The proposed model for this narrative example of the Interim Procedures is AQ40, a hypothetical dispersion model. The computer code for AQ40 embodies the features of several publicly available Gaussian dispersion models. Because of resource constraints for preparing this example, the proposed model description and the technical model comparisons have been foreshortened. For this example, familiarity with the features of the MPTER model is assumed, and the technical comparison presents only the key technical differences between AQ40 and MPTER. For an actual application of the Interim Procedures, a complete technical description of the proposed model should be prepared.

Preliminary estimates of the SO_2 impact of the Clifty Creek plant were obtained using EPA screening techniques as recommended in the Interim Procedures. These estimates indicate that maximum concentrations occur within 10 kilometers of the Clifty Creek generating station. Refined modeling using the proposed and reference models AQ40 and MPTER, respectively, with 1975 hourly meteorological data has also been done. On the basis of the AQ40 modeling results, the 3- and 24-hour average maximum SO_2 concentrations would be expected to occur approximately 3 kilometers south of the plant. MPTER predicts that both the 3- and 24-hour average maximum SO_2 impacts will occur approximately 7 kilometers northeast of the Clifty Creek station. Results of this preliminary modeling are to be used in designing an appropriate performance evaluation data network by indicating potential maximum impact areas and are useful in designing the statistical model comparison methodology required by the Interim Procedures. Once the preliminary ambient estimates have been made, the next step in applying the Interim Procedures For Evaluting Air Quality Models is to perform a technical comparison of the proposed and reference models. The technical

comparison of the proposed and reference models should then be performed following the methodology set forth in the Workbook For Comparison of Air Quality Models.[3] The purpose of the technical model comparison is to determine which model would be expected to predict more accurately concentrations for the source being considered. If results of the statistical performance comparisons, carried out in a subsequent step, are inconclusive, the results of the technical model comparison can serve as the bias for determining the acceptability of the proposed model.

The important technical differences between AQ40 and MPTER are:

(a) Terrain considerations. MPTER simulates the effect of terrain by subtracting the full terrain height from the effective plume height. AQ40 uses full terrain subtraction from the effective plume height for stable atmospheric mixing conditions and half terrain height subtraction for neutral and unstable meteorology.

(b) Dispersion coefficients. MPTER uses the Pasquill-Gifford horizontal and vertical dispersion coefficients and six stability classes. AQ40 uses the rural ASME[4] horizontal and vertical dispersion coefficients and five stability classes (one stable class).

(c) Stack tip downwash. MPTER, as run for this example evaluation, does not invoke this option. AQ40 does simulate this phenomenon.

(d) Plume rise. MPTER uses the final Briggs' plume rise approximation. AQ40 uses the transitional or distance-dependent Briggs' plume rise formulation.

(e) Buoyancy induced dispersion. MPTER does not enhance dispersion due to buoyantly rising plumes, but AQ40 does employ this option.

(f) Wind profile. MPTER and AQ40 both use a power law for adjusting wind speed with height, but use different coefficients, as presented in

the following table. AQ40 uses the predicted wind speed at final plume height in the denominator of the Gaussian dispersion equation. MPTER uses the predicted wind speed at stack height in the Gaussian equation.

Stability	1	2	3	4	5	6
MPTER	.1	.15	.2	.25	.3	.3
AQ40	.10	.11	.12	.15	.20	none

(g) Mixing height. With both MPTER and AQ40, the mixing height rises and falls to maintain a constant height above local terrain. For MPTER, however, plumes rising above the mixing height have no ground-level impact and plumes below the mixing height are fully reflected. With AQ40, on the other hand, unlimited mixing heights are used for stable atmospheric conditions, while a partial plume penetration algorithm is employed for nonstable conditions.

In an actual model evaluation, a complete technical model comparison using the "Workbook" procedures would be carried out and submitted to the control agency for review and agreement that both the proposed and reference models are appropriate for the regulatory application at hand.

3.0 Model Evaluation Protocol

As previously stated, the two principal regulatory purposes that this evaluation protocol addresses are the following:

° Compliance with National Ambient Air Quality Standards (NAAQS) for sulfur dioxide (SO_2) for 3-hour and 24-hour averaging times.

° Assessment of plant SO_2 impact on a hypothetical Class I Prevention of Significant Deterioration (PSD) region located 15 kilometers northeast of the plant (3-hour and 24-hour averaging times in the vicinity of the Bacon Ridge Site.)

The performance of the proposed and reference models, AQ40 and MPTER, respectively, will be compared based upon each model's ability to simulate air quality impacts measured on a monitoring network of six SO_2 stations in the vicinity of the Clifty Creek generating station. The period of record for the concurrent air quality, meteorological and source data proposed for these evaluations is January 1 through December 31, 1976.

Since the projected impact areas are different for each regulatory purpose, the performance of the models for NAAQS and PSD applications will be assessed independently. The performance of the models for NAAQS will be judged based upon the data from the entire six station network for 1-, 3- and 24-hour averaging times. Model performance for the PSD application will emphasize data from the Bacon Ridge Station within the hypothetical Class I region. It is possible that different models may be selected as being most appropriate for each of the above issues.

3.1 NAAQS Attainment

Three performance evaluation objectives have been established which are important with respect to this primary regulatory purpose.

The first-order objective is to test the ability of the models to predict successfully the highest concentrations for use in the regulatory decision-making process. It is recognized that the single-point prediction of the highest, second-highest concentration is statistically unmeaningful; therefore, performance measures in this group also include analysis of the uppermost predicted and observed concentrations for the data period of record.

The second-order objective is to test the ability of the models to predict successfully the entire domain of concentrations.

The third-order objective is to test the ability of the models to predict successfully the spatial and temporal patterns of concentrations. Tables 1 through 3 summarize the model comparison protocol for the NAAQS analysis. The tables describe the evaluation data sets, the performance measures, bases for calculating confidence intervals, averaging times to which the performance measures will be applied, and the point assignments for each measure that will be used to score and compare the predictive abilities of the two models.

The performance measures listed in Tables 1 through 3 were selected to reflect the spirit of the American Meteorological Society (AMS)[5] recommendations. The listed performance measures are specifically those required to test the ability of the models to meet the model performance objectives stated above. The AMS recommendations define statistical procedures for comparing model predictions with observed concentration values. In this example, two models are being compared based on how each performs against the same set of observations. This three-way comparison (proposed model vs. reference model vs. observations) poses a formidable

TABLE 1

MODEL COMPARISON PROTOCOL FOR NAAQS ANALYSIS
First-Order Objective: Predict Highest Concentrations

Data Set (Subset)	Pairing		Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Maximum Points	Rationale
	Space	Time					
Highest concentration (predicted and observed) in the monitoring network. (A-3)	No	No	Direct comparison of values only.	None applicable	3 24	20 20	Measure of the models' ability to predict the highest concentration value; however, single-point predictions are statistically unmeaningful, so relatively low points have been assigned.
Highest, second-highest concentration (predicted and observed) in the monitoring network. (A-3)	No	No	Direct comparison of values only.	None applicable	3 24	30 30	NAAQS regulations are based on the highest second-highest value, however, single-point predictions are statistically unmeaningful, so relatively low points have been assigned.
Uppermost 5% (or 25) observed vs. predicted concentrations. (A-4a)	No	No	Bias	Two sample t-test	1 3 24	25 15 15	Measures of models' ability to predict high concentration values without regard to time or location. Expected to be more stable statistical measures than single-point predictions, and therefore are assigned more points than the single-point measures above.
	No	No	Variance	F-test	1 3 24	10 5 5	
	No	No	Goodness of fit	Kolmogorov-Smirnov	1 3 24	25 15 15	

TABLE 1
(Continued)

MODEL COMPARISON PROTOCOL FOR NAAQS ANALYSIS
First-Order Objective: Predict Highest Concentrations

Data Set (Subset)	Pairing		Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Maximum Points	Rationale
	Space	Time					
Observed vs. predicted maximum concentration in the monitoring network by time event. (A-1)	No	Yes	Bias	One sample t-test	1	15	Measures of model's ability to predict high concentrations in the network on an event-by-event basis (temporal pattern of maxima).
					3	10	
					24	10	
	No	Yes	Variance of residual	x^2- test	1	10	
					3	5	
					24	5	
	No	No	Goodness of fit	Kolmogorov-Smirnov	1	15	
					3	10	
					24	10	
Uppermost 5% (or 25) observed vs. predicted concentrations for the particular station exhibiting the greatest number of high observed concentrations. This station will be identified as the station having the greatest number of concentrations within the highest 5% (or 25) of concentrations observed over all stations. (A-4b)	Yes	No	Bias	Two sample t-test	1	25	Measures of models' ability to predict high concentration values at the receptor exhibiting the greatest number of high-concentration impacts.
					3	15	
					24	15	
	Yes	No	Variance	F-test	1	10	
					3	5	
					24	5	
	Yes	No	Goodness of fit	Kolmogorov-Smirnov	1	25	
					3	15	
					24	15	

TABLE 1
(Continued)

MODEL COMPARISON PROTOCOL FOR NAAQS ANALYSIS
First-Order Objective: Predict Highest Concentrations

Data Set (Subset)	Pairing		Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Maximum Points	Rationale
	Space	Time					
Observed vs. predicted maximum concentrations in the monitoring network for the stability condition associated with the greatest number of observations in the uppermost 5% (or 25) observed concentrations. (A-5)	No	No	Bias	Two sample t-test	1	20	Measures of models' ability to predict maximum concentrations for the stability conditions associated with the highest observed concentrations.
	No	No	Variance	F-test	1	10	
	No	No	Goodness of fit	Kolmogorov-Smirnov	1	20	

TABLE 2

MODEL COMPARISON PROTOCOL FOR NAAQS ANALYSIS

Second-Order Objective: Predict the Domain of Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Maximum Points	Rationale
All observed vs. predicted concentration at each station. (B-1)	Yes	Yes	Bias	One sample t-test	1	25	Measures of model's ability to predict concentration values at each receptor. Since there are six stations, one-sixth of the maximum points will be awarded for the comparisons of each station.
					3	15	
					24	15	
	Yes	Yes	Variance of residual	x^2- test	1	10	
					3	5	
					24	5	
	Yes	No	Goodness of fit	Kolmogorov-Smirnov	1	20	
					3	10	
					24	10	
Observed vs. predicted concentrations for entire data set. (B-3)	Yes	Yes	Bias	One sample t-test	1	25	Measures of models' ability to predict concentrations at particular places and times (i.e., domain of concentrations). This is a very strict statistical test of model performance.
					3	15	
					24	15	
	Yes	Yes	Variance of residual	x^2- test	1	10	
					3	5	
					24	5	
	No	No	Goodness of fit	Kolmogorov-Smirnov	1	20	
					3	10	
					24	10	
Observed vs. predicted concentrations for unstable (A, B); meteorological conditions (C); neutral meteorological conditions (D); and stable meteorological conditions (E, F). (B-4)	Yes	Yes	Bias	One sample t-test	1	20	Measures of models' ability to predict concentrations at specified times and places for various atmospheric stability classes. Since there are four stability groups, one-fourth of the maximum available points will be awarded for the comparison in each stability class group.
	Yes	Yes	Variance of residual	x^2- test	1	10	
	No	No	Goodness of fit	Kolmogorov-Smirnov	1	15	

TABLE 2
(Continued)

MODEL COMPARISON PROTOCOL FOR NAAQS ANALYSIS
Second-Order Objective: Predict the Domain of Concentrations

Data Set (Subset)	Pairing		Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Maximum Points	Rationale
	Space	Time					
Observed vs. predicted concentrations for the following windspeed groups (<2.5 m/sec); (2.5 - 5 m/sec); and (>5 m/sec). (B-4)	Yes	Yes	Bias	One sample t-test	1	10	Measure of models' ability to predict concentrations at specified times and places for low and high windspeed classes. Since there are three windspeed groups, one-third of the maximum available points will be awarded for the comparison in each windspeed class.
	Yes	Yes	Variance of residual	x^2- test	1	5	
	No	No	Goodness of fit	Kolmogorov-Smirnov	1	10	

TABLE 3

MODEL COMPARISON PROTOCOL FOR NAAQS ANALYSIS
Third-Order Objective: Predict the Pattern (Spatial and Temporal) of Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Maximum Points	Rationale
All observed vs. predicted concentration at each station. (B-1)	Yes	Yes	Correlation coefficient	Fisher-Z	1 3 24	25 15 15	Measures of association between predicted and observed concentration values at each receptor (a measure of the ability of the model to predict the temporal pattern). Since there are six stations, one-sixth of the maximum points will be awarded for the comparison of each station.
Observed vs. predicted maximum concentration in the monitoring network by time event. (A-1)	No	Yes	Correlation coefficient	Fisher-Z	1 3 24	25 15 15	Measures of association between predicted and observed high concentrations in the network on an event-by-event basis (temporal pattern of maxima).
Observed vs. predicted concentrations for entire data set. (B-3)	Yes	Yes	Correlation coefficient	Fisher-Z	1 3 24	40 25 25	Measure of association between predicted and observed concentrations, on an hour-by-hour basis.

problem for which appropriate statistical methods have not yet been devised. The procedures described below for comparing models provide a decision making framework based upon standard statistical measures.

The first three columns in Tables 1 through 3 describe the data sets being used. The letter and number code in parentheses in the first column are included for cross reference to a numbering system recently prepared by EPA (See Table 3.2 Interim Procedures for Evaluating Air Quality Models[1]). The fourth and fifth columns specify the performance measure being addressed and the statistical method that will be used to assign the 95 percent confidence band about each performance measure. The sixth column lists the averaging times to which each performance measure will be applied. The seventh column lists the points assigned to each performance measure for scoring model performance. The final column briefly discusses each of the performance measures, providing a rational for using each data subset and group of performance measures.

Tables 1 through 3 contain 67 performance measures designed to test the relative abilities of AQ40 and MPTER to meet the three evaluation objectives. In assigning points to each performance measure, an attempt was made to balance the regulatory importance, statistical significance and scientific value of each performance measure. A total of 1,000 possible points has been divided among the three model evaluation objectives. In recognition of the regulatory importance of the first-order model evaluation objective, one-half of the total available points (500) have been assigned to the set of performance measures grouped under that objective, that is, the ability of the models to predict the highest concentration values. The second performance objective, prediction of

the domain of concentration, has been assigned 300 points, and the third performance objective, prediction of the pattern of concentrations, has been allotted the remaining 200 points.

Four types of performance measures and associated statistical tests are being used to judge model performance. The performance measures (and associated statistical tests) are the absolute value of the bias (t-test), variance (F-test and X^2), goodness of fit (Kolmogorov-Smirnov test), and correlation coefficient (Fisher-z test). Errors of magnitude of prediction are considered to be more critical than errors of scatter of prediction, therefore measures of bias and goodness of fit (which test magnitude errors) have been alloted more points than measures of variance and correlation. Since the basic prediction time step of both MPTER and AQ40 is 1 hour, the 1-hour averaging time measures have received more points than the regulatory averaging times (3- and 24-hours). This is done following the recommendations of the AMS Workshop on Dispersion Model Performance.

Performance measures and confidence intervals for each performance measure will be calculated for both MPTER and AQ40 for the averaging times indicated in Tables 1 through 3. The performance of the models will be compared, performance measure by performance measure and averaging time by averaging time. If the performance of the two models is significantly different statistically (that is, the 95 percent confidence interval for either model does not include the value of the performance measure for the other model) the points indicated in Tables 1 through 3 will be awarded to the model that calculates closest to the observed value. Positive points are accumulated for each performance measure if the proposed model performs better; negative points are accumulated if the

reference model shows superior performance. For goodness of fit all the points (plus or minus) will be awarded based upon which model has better statistical performance.

If, for the two models, the 95 percent confidence intervals for the absolute value of the bias, variance, or correlation measures do contain the values of the performance measures for both models, the non-overlapping confidence intervals for those measures will be calculated, and the corresponding percentage of the maximum available points will be assigned. For example, if the 95 percent confidence intervals for the bias of the two models overlap each other's mean bias (see Figure 2), the confidence intervals of measures for both models will be "tightened" (See Appendix C) until the two biases are mutually different statistically at some level of significance, as in Figure 3. To illustrate, assume that bias is a 10-point measure and assume the biases become statistically distinct at the 90 percent confidence level; then nine (90 percent) of the possible 10 points would be awarded to the model that better predicts the bias. Only integer points will be awarded as fractions will be rounded. (Although it is recognized that this methodology may not be ideal in the strictest statistical sense, it is acceptable for example purposes and is easy to apply. Others may wish to propose another methodology for scoring.)

Following the completion of all the performance measure comparisons, the points awarded will be totalled. If the grand total is $\geq$ +100 points, the proposed model will be deemed more suitable for assessing the plant impact for the appropriate NAAQS averaging time. If the grand total is between -100 and +100 points, no decisive conclusion may be reached regarding the superiority of either model and further analysis would be considered (for example, technical comparisons or further evaluations

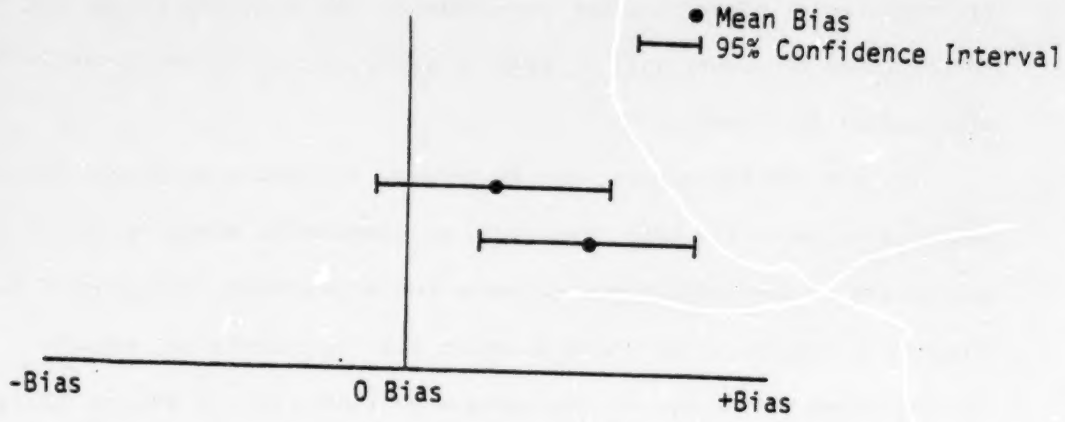


Figure 2. Example of Overlapping 95% Confidence Intervals on Bias for Two Models

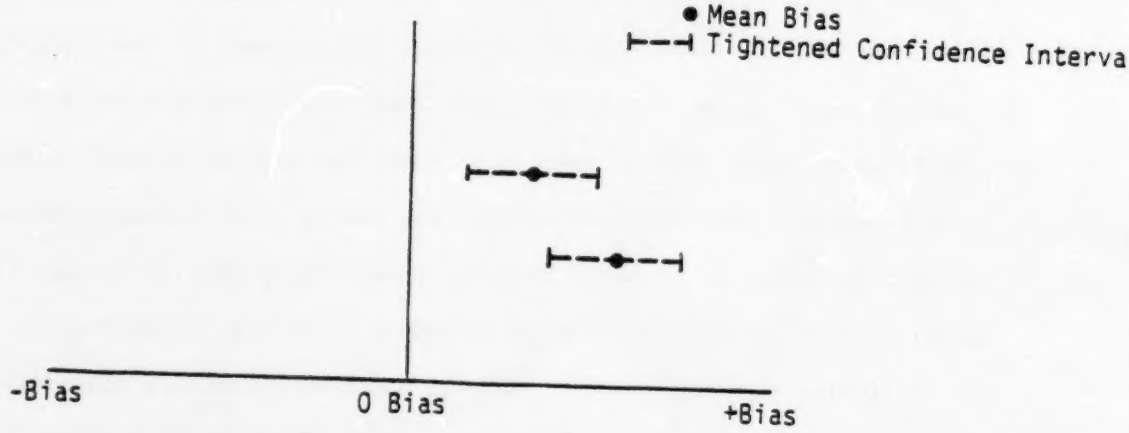


Figure 3. Example of "Tightened" Confidence Intervals to Result in Non-Overlapping Biases

with more data - see Section 2.) If the grand total is $\leq$ -100 points, the reference model will be judged to have the better performance.

3.2 PSD Analysis

As with the NAAQS analysis, three performance objectives have been established to assess the performance of the two models. The major difference between the two analyses results from the fact that only one station is available in the Class I PSD area, which reduces the number of data sets used in the PSD portion of the model evaluation.

Table 4 summarizes the first-order objectives and associated performance measures designed to assess the ability of the models to predict concentrations as required for a Class I PSD analysis. That is, the data subsets and performance measures in Table 4 evaluate the ability of the models to predict highest impacts within the hypothetical Class I area described previously. The second- and third-order objectives and performance measures for evaluating the models in this PSD application are identical to those presented previously in Tables 2 and 3 for NAAQS analysis. Again, one-half the points have been allotted to the group of performance measures included under the first-order objective (500 points), that is, testing the ability of the models to predict highest concentrations at the station located within the PSD Class I area. The remaining points are assigned to the second- and third-order objectives in the same manner as was done for the NAAQS model evaluation. The performance of the models for the PSD application will be compared after summing the points that each model scores in Tables 2, 3, and 4. Awarding of points and identification of the model with superior performance are accomplished in a manner identical to the method used for the NAAQS analysis.

TABLE 4

MODEL COMPARISON PROTOCOL FOR PSD ANALYSIS
First Order Objective: Predict Highest Concentrations in PSD Area

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Maximum Points	Rationale
Highest concentration (predicted and observed) at the station within the PSD, Class I area. (A-3)	Yes	No	Direct comparison of values only	None applicable	3 24	50 50	Measure of the models' ability to predict the highest concentration value; however, single-point predictions are statistically unmeaningful, so relatively low points have been assigned.
Highest, second-highest concentration (predicted and observed) for the station within the Class I area.	Yes	No	Direct comparison of values only	None applicable	3 24	60 60	PSD regulations are based on the highest second-highest value; however, single-point predictions are statistically unmeaningful, so relatively low points have been assigned.
Uppermost 5% (or 25) observed and predicted concentrations for the station within the Class I area. (A-4a)	Yes	No	Bias	Two sample t-test	1 3 24	50 30 30	Measures of models' ability to predict high concentration values without regard to time at the station within the Class I region. These measures receive more points than the single-point measures above because they are statistically more stable.
	Yes	No	Variance	F-test	1 3 24	30 15 15	
	Yes	No	Goodness of fit	Kolmogorov-Smirnov	1 3 24	50 30 30	

The above methodology will result in the objective selection of the reference or proposed model for each of the regulatory situations of concern.

4.0 Field Measurements

Following EPA concurrence that the protocol for model performance evaluation is technically sound, the field measurements program is undertaken. The purpose of the field measurements program is to generate a data base to be used for the comparative model performance evaluations. The field program design is based upon the results of the preliminary modeling, as discussed in Section 2, and the requirements of the protocol, as discussed in Section 3. For this example, resource constraints precluded the design of a data acquisition network and collection of the requisite field data. Instead a historical data base is used and a hypothetical regulatory problem is constructed around that data base with the primary goal to illustrate the use of the statistical performance evaluation methodology described in the Interim Procedures. A real regulatory problem would require an in-depth analysis of the data requirements for a comparative model evaluation.

Preceding page blank

5.0 Performance Evaluation Results and Model Selection

After completion of the field measurements program, the data collected are used for the comparative model performance evaluation. The performance evaluation follows the plan presented in the protocol. Once the results of the performance evaluation are compiled, the decision is made whether to accept or reject the proposed model based upon the objective scoring scheme presented in the protocol. A report containing the results of the evaluations, the results of the comparative model scoring, and the decision whether or not to accept the proposed model is submitted to the control agency for review. It is essential to calculate the statistical performance measures and apply the decision criteria exactly as specified in the preplanned protocol. Adherence to the protocol ensures that the decision is completely objective.

For this example evaluation, it is assumed that the control agency approved the performance measures and scoring scheme proposed in the protocol as presented in Section 3. Tables 5 through 8 present the results of the example model evaluations called for in Tables 1 through 4. The first three columns of each table list the data sets being compared and indicate whether or not the observed and predicted concentrations are paired in time or space. The next two columns list the statistical performance measures and the statistical bases for calculating confidence intervals for use in scoring performance. The sixth column gives the averaging time for which each comparison has been made. The next three columns list the actual performance of MPTER and AQ40, and, where applicable, the level of significance at which the models demonstrate statistically different performance. Where the variance ratio is utilized

Preceding page blank

TABLE 5

MODEL COMPARISON RESULTS FOR NAAQS ANALYSIS
First-Order Objective: Predict Highest Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: MPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different[a]	Maximum Points	Points Awarded
Highest concentration (predicted and observed) in the monitoring network. (A-3)	No	No	Direct comparison of values only.	None applicable	3	-232.3	-323.4	NC	20	-20
			C_o max-C_p max		24	-67.5	2.5	NC	20	+20
Highest, second-high concentration (predicted and observed) in the monitoring network.	No	No	Direct comparison of values only.	None applicable	3	-246.5	-134.3	NC	30	+30
			C_o 2 max-C_p 2 max		24	-18.0	-14.7	NC	30	+30
Uppermost 5% (or 25) observed vs. predicted concentrations. (A-4a)	No	No	Bias	Two sample t-test	1	-280.4	-246.0	40%	25	+10
			C_o-C_p		3	-76.6	-50.9	40%	15	+6
					24	-1.4	23.3	95%	15	-15
	No	No	Variance ratio	F-test						
			(Observed variance)	= (10967)	1	0.49	0.22	95%	10	-10
				(7950)	3	0.63	0.47	75%	5	-4
				(380)	24	2.99	0.59	95%	5	+5
	NO	No	Goodness of fit	Kolmogorov-Smirnov	1	.80	.72	NC	25	+25
					3	.56	.36	NC	15	+15
					24	.20	-.72	NC	15	-15

[a] NC = not calculated

TABLE 5
(Continued)

MODEL COMPARISON RESULTS FOR NAAQS ANALYSIS
First-Order Objective: Predict Highest Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: MPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different[a]	Maximum Points	Points Awarded
Observed vs. predicted maximum concentration in the monitoring network by time event. (A-1)	No	Yes	Bias $C_o - C_p$	One sample t-test	1	45.9	50.0	40%	15	-6
					3	39.5	42.4	40%	10	-4
					24	18.0	20.3	60%	10	-6
	No	Yes	Variance of residual	x^2- test	1	39877	34695	95%	10	+10
					3	16485	14520	95%	5	+5
					24	1314	1125	90%	5	+4
	No	No	Goodness of fit	Kolmogorov-Smirnov	1	-.20	-.60	NC	15	-15
					3	-.17	-.16	NC	10	+10
					24	.79	.09	NC	10	+10
Uppermost 5% (or 25) observed vs. predicted concentrations for the particular station exhibiting the greatest number of high observed concentrations. This station is identified as the station having the greatest number of concentrations within the highest 5% (or 25) of concentrations observed over all stations. (A-4b) That station is Rykers Ridge.	Yes	No	Bias	Two sample t-test	1	-67.2	21.1	80%	25	+20
					3	46.2	103.2	90%	15	-14
					24	27.4	36.6	80%	15	-12
	Yes	No	Variance ratio (Observed variance)	F-test = (31732)	1	0.53	1.07	95%	10	+10
				(9828)	3	0.63	0.82	60%	5	+3
				(341)	24	0.60	1.62	95%	5	-5
	Yes	No	Goodness of fit	Kolmogorov-Smirnov	1	0.28	-0.20	NC	25	+25
					3	-0.48	-0.72	NC	15	-15
					24	-0.64	-0.84	NC	15	-15

[a] NC = not calculated

TABLE 5
(Continued)

MODEL COMPARISON RESULTS FOR NAAQS ANALYSIS
First-Order Objective: Predict Highest Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: HPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different*	Maximum Points	Points Awarded
Observed vs. predicted maximum concentrations in the monitoring network for the stability condition associated with the greatest number of observations in the uppermost 5% (or 25) observed concentrations. (A-5) That stability is D (4).	No	No	Bias $C_o - C_p$	Two sample t-test	1	166.4	303.5	95%	20	-20
			Variance ratio (Observed variance) = (10280)	F-test	1	3.55	0.90	95%	10	+10
			Goodness of fit	Kolmogorov-Smirnov	1	-0.76	-0.92	NC	20	-20
								Subtotal	±500	+52

* NC = not calculated

TABLE 6

MODEL COMPARISON RESULTS FOR NAAQS ANALYSIS
Second-Order Objective: Predict the Domain of Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: HPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different*	Maximum Points	Points Awarded
All observed vs. predicted concentration at each station. (B-1)	Yes	Yes	Bias	One sample t-test	1				25	+1
Bacon Ridge			C_o-C_p							
Rykers Ridge						42.6	56.6	90%		
North Madison						58.9	59.6	<20%		
Hebron Church						11.8	9.1	20%		
Liberty Ridge						11.0	27.8	90%		
Canip Creek						55.0	34.6	95%		
						92.1	79.9	90%		
Bacon Ridge										
Rykers Ridge					3	34.1	45.9	90%	15	0
North Madison						49.3	48.6	<20%		
Hebron Church						7.6	5.6	<20%		
Liberty Ridge						8.7	22.7	90%		
Canip Creek						48.1	22.5	95%		
						74.4	65.1	80%		
Bacon Ridge										
Rykers Ridge					24	13.7	17.4	80%	15	0
North Madison						16.8	16.9	<20%		
Hebron Church						5.8	5.4	<20%		
Liberty Ridge						5.2	9.5	90%		
Canip Creek						12.6	8.7	95%		
						23.3	20.9	60%		
	Yes	Yes	Variance of residual	x^2- test	1				10	+5
Bacon Ridge										
Rykers Ridge						33402	28831	95%		
North Madison						35717	30682	95%		
Hebron Church						50767	42164	95%		
Liberty Ridge						41163	34193	95%		
Canip Creek						16558	28528	95%		
						20138	19900	<20%		
Bacon Ridge										
Rykers Ridge					3	17567	14360	95%	5	+2
North Madison						14856	12058	95%		
Hebron Church						21641	20718	20%		
Liberty Ridge						18280	16282	90%		
Canip Creek						6357	9066	95%		
						8322	7337	80%		
Bacon Ridge										
Rykers Ridge					24	1251	976	95%	5	+2
North Madison						675	582	80%		
Hebron Church						1411	1157	90%		
Liberty Ridge						1094	981	60%		
Canip Creek						187	255	95%		
						588	529	60%		

* NC = not calculated

TABLE 6
(Continued)

MODEL COMPARISON RESULTS FOR NAAQS ANALYSIS
Second-Order Objective: Predict the Domain of Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: MPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different*	Maximum Points	Points Awarded
	Yes	No	Goodness of fit	Kolmogorov-Smirnov	1				20	+20
Bacon Ridge						-.60	-.21	NC		
Rykers Ridge						-.75	-.25	NC		
North Madison						-.60	-.22	NC		
Hebron Church						-.35	-.12	NC		
Liberty Ridge						-.91	-.36	NC		
Canip Creek						-.89	-.21	NC		
Bacon Ridge					3	-.16	-.14	NC	10	+5
Rykers Ridge						-.22	-.19	NC		
North Madison						-.23	-.19	NC		
Hebron Church						.09	.09	NC		
Liberty Ridge						-.29	-.30	NC		
Canip Creek						.21	.17	NC		
Bacon Ridge					24	.13	.16	NC	10	-2
Rykers Ridge						-.16	.14	NC		
North Madison						-.23	.17	NC		
Hebron Church						.13	.14	NC		
Liberty Ridge						-.20	-.24	NC		
Canip Creek						-.13	.13	NC		
Observed vs. predicted concentrations for entire data set. (B-3)	Yes	Yes	Bias C_o-C_p	One sample t-test	1	41.8	44.4	40%	25	-10
					3	33.5	35.9	40%	15	-6
					24	12.7	13.4	40%	15	-6
	Yes	Yes	Variance of residual	x^2- test	1	36082	31871	95%	10	+10
					3	16400	14520	95%	5	+5
					24	962	815	95%	5	+5
	No	No	Goodness of fit	Kolmogorov-Smirnov	1	-.63	-.22	NC	20	+20
					3	-.19	-.16	NC	10	+10
					24	-.17	.14	NC	10	+10
Observed vs. predicted concentrations for unstable (A, B); meteorological conditions (C); neutral meteorological conditions (D); and stable meteorological conditions (E, F). (B-4)	Yes	Yes	Bias C_o-C_p	One sample t-test	1				20	-2
A, B						-28.0	-29.4	<20%		
C						-6.9	-4.4	<20%		
D						70.6	67.0	60%		
E, F						38.9	63.[illegible]	[illegible]		

TABLE 6
(Continued)

MODEL COMPARISON RESULTS FOR NAAQS ANALYSIS
Second-Order Objective: Predict the Domain of Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: MPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different*	Maximum Points	Points Awarded
	Yes	Yes	Variance of residual	χ^2- test	1				10	+10
A, B						91880	82161	90%		
C						70489	63431	95%		
D						22088	19432	95%		
E, F						10366	4084	95%		
	No	No	Goodness of fit	Kolmogorov-Smirnov	1				15	+15
A, B						-.43	-.19	NC		
C						-.46	-.19	NC		
D						-.66	-.13	NC		
E, F						-.81	-.26	NC		
Observed vs. predicted concentrations for the following wind-speed groups (<2.5 m/sec) (<2.5 - 5 m/sec); and (>5 m/sec). (B-4)	Yes	Yes	Bias C_o-C_p	One sample t-test	1				10	-3
<2.5 m/sec						38.9	18.6	95%		
2.5 - 5 m/sec						46.6	59.0	95%		
>5 m/sec						15.1	50.7	95%		
	Yes	Yes	Variance of residual	χ^2- test	1				5	+2
<2.5 m/sec						38587	45834	95%		
2.5 - 5 m/sec						33071	23928	95%		
>5 m/sec						48609	24567	95%		
	No	No	Goodness of fit	Kolmogorov-Smirnov	1				10	+10
<2.5 m/sec						-.71	-.27	NC		
2.5 - 5 m/sec						-.61	-.19	NC		
>5 m/sec						-.35	.08	NC		
								Subtotal	±300	+103

* NC = not calculated

TABLE 7

MODEL COMPARISON RESULTS FOR NAAQS ANALYSIS
Third-Order Objective: Predict the Pattern (Spatial and Temporal) of Concentrations

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: MPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different[a]	Maximum Points	Points Awarded
All observed vs. predicted concentration at each station. (B-1)	Yes	Yes	Correlation Coefficient	Fisher-Z	1				25	-8
Bacon Ridge						-.11	-.10	20%		
Rykers Ridge						-.01	-.01	<20%		
North Madison						-.01	-.05	60%		
Hebron Church						-.10	-.13	60%		
Liberty Ridge						.14	.10	60%		
Canip Creek						.08	.06	40%		
Bacon Ridge					3	-.15	-.13	20%	15	+2
Rykers Ridge						.05	.07	20%		
North Madison						-.03	-.06	40%		
Hebron Church						-.07	-.06	<20%		
Liberty Ridge						.29	.29	<20%		
Canip Creek						.16	.21	60%		
Bacon Ridge					24	-.14	-.18	40%	15	+2
Rykers Ridge						.34	.37	20%		
North Madison						.17	.16	<20%		
Hebron Church						.49	.46	40%		
Liberty Ridge						.52	.49	20%		
Canip Creek						.16	.28	80%		
Observed vs. predicted maximum concentration in the monitoring network by time event. (A-1)	No	Yes	Correlation Coefficient	Fisher-Z	1	.17	.13	95%	25	-25
					3	.24	.22	40%	15	-6
					24	.49	.44	60%	15	-9
Observed vs. predicted concentrations for entire data set. (B-3)	Yes	Yes	Correlation Coefficient	Fisher-Z	1	-.04	-.05	40%	40	-16
					3	-.02	-.01	20%	25	+5
					24	.29	.29	<20%	25	0
								Subtotal	±200	-55
								Grand total points for NAAQS analysis	±1000	+100

[a] NC = not calculated

TABLE 8

MODEL COMPARISON RESULTS FOR PSD ANALYSIS
First-Order Objective: Predict Highest Concentrations in PSD Area

Data Set (Subset)	Pairing: Space	Pairing: Time	Performance Measures	Statistical Basis For Confidence Interval	Averaging Time	Performance: MPTER	Performance: AQ40	Level of Significance at which Model Performances are Statistically Different[a]	Maximum Points	Points Awarded
Highest concentration (predicted and observed) at the station within the PSD, Class I area. (A-3) That station is Bacon Ridge.	Yes	No	Direct comparison of values only $C_{o\ max}$-$C_{p\ max}$	None applicable	3	67.7	174.2	NC	50	-50
					24	6.3	49.5	NC	50	-50
Second-highest concentration (predicted and observed) for the station within the Class I area. That station is Bacon Ridge.	Yes	No	Direct comparison of values only $C_{o\ 2\ max}$-$C_{p\ 2\ max}$	None applicable	3	66.7	145.3	NC	60	-60
					24	14.3	58.4	NC	60	-60
Uppermost 5% (or 25) observed and predicted concentrations for the station within the Class I area. That station is Bacon Ridge.	Yes	No	Bias C_o-C_p	One sample t-test	1	-2.0	70.6	90%	50	-45
					3	31.0	108.0	95%	30	-30
					24	20.8	40.0	95%	30	-30
	Yes	No	Variance ratio (Observed variance)	F-test = (15446)	1	0.93	0.28	95%	30	-30
				(8737)	3	0.80	1.37	90%	15	-14
				(363)	24	0.42	1.02	95%	15	+15
	Yes	No	Goodness of fit	Kolmogorov-Smirnov	1	.20	-.60	NC	50	-50
					3	-.44	-.68	NC	30	-30
					24	.64	-.76	NC	30	-30
								Subtotal	±500	-464
								Grand total points for PSD analysis	±1000	-416

[a] NC = not calculated

as a performance measure, the variance of the observed concentrations is also provided to aid in data interpretation. The last two columns list the maximum points that may be awarded for each performance measure/averaging time combination, and the points that actually were awarded. At the end of each table the subtotal of points awarded is indicated. At the end of Tables 7 and 8, the grand total points for the model performance comparisons are presented for the NAAQS application and the PSD application, respectively.

5.1 Results for Model Performance Comparisons in the NAAQS Analysis

The grand total points awarded for the model performance comparison in the NAAQS analysis is +100 points out of a possible $\pm$1000 points. As set forth in the protocol, a score from +100 to +1000 points results in the acceptance of AQ40, the proposed model, over MPTER, the reference model, for NAAQS regulatory applications at the Clifty Creek station. With a score of +100, AQ40 has attained the minimum score required for acceptance over the reference model.

Inspection of Table 5 reveals that in the NAAQS application AQ40 better predicted the highest, second-highest observed SO_2 concentrations for both 3- and 24-hour averaging times and showed slightly better performance overall for the first performance evaluation objective, predicting the highest concentrations (+52 points out of a possible $\pm$500 points). AQ40 also outperformed MPTER on the second objective, predicting the domain of concentrations, by scoring +103 out of a possible $\pm$300 points (see Table 6). MPTER, however, scored better (with -55 points out of $\pm$200 possible points) for the third objective which tests the ability of the models to match spatial and temporal patterns of concentration

(see Table 7). Based upon the grand total points awarded, AQ40 is accepted as suitable for the regulatory analyses of 3- and 24-hour average SO_2 impacts from the Clifty Creek generating station in relation to NAAQS requirements.

5.2 Results for Model Performance Comparisons in the PSD Analysis

The grand total points awarded for the model performance comparison in the PSD analysis is -416 points out of a possible $\pm$1000 points, as shown in Table 8. As set forth in the protocol, a score from -100 to -1000 points results in the rejection of the proposed model. The score of -416 for the PSD comparative performance evaluation indicates a decisive margin in favor of the reference model MPTER over the proposed model AQ40.

Inspection of Table 8 shows that MPTER outperformed AQ40 for the first-order performance evaluation objective of predicting the highest concentrations within the PSD region. MPTER scored -464 points out of a possible $\pm$500 points for this first-order objective, and more accurately predicted both the 3- and 24-hour average highest, second-highest concentrations as monitored within the PSD region.

As stipulated by the protocol, the model performance measures and scoring scheme used for the second- and third-order PSD evaluation objectives, predicting the domain of concentrations, respectively, are identical to the performance measures and scoring scheme used for the NAAQS model comparisons. The performance results and points awarded for these comparisons were presented previously in Tables 6 and 7. Recall that AQ40 scored +103 points out of $\pm$300 possible points for the second-order objective, and that MPTER scored -55 points of $\pm$200 possible points

for the third-order objective. Based upon the grand total points awarded, AQ40 is rejected as being suitable for the regulatory analyses of 3- and 24-hour average SO_2 impacts from the Clifty Creek generating station within the hypothetical PSD Class I area.

6.0 Summary

This narrative example of the Interim Procedures For Evaluating Air Quality Models illustrates the analytical steps necessary to judge the acceptability of a proposed, non-Guideline model for a specific regulatory application. The statistical performance measures and scoring methodology used in the narrative example have been selected for this hypothetical application and set forth in the preplanned protocol. In actual applications of the Interim Procedures, the performance evaluation methodology must be designed to meet the specitic objectives of the intended regulatory use of the proposed model. It is especially important that close liaison be maintained with the control agency throughout the model evaluation process to ensure agreement on the objectivity of the model comparison results.

7.0 References

1. Environmental Protection Agency. "Interim Procedures For Evaluating Air Quality Models." Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina, August 1981.

2. Environmental Protection Agency. "Guideline On Air Quality Models." EPA 450/2-78-027. Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina, April 1981.

3. Environmental Protection Agency. "Workbook For Comparison Of Air Quality Models." EPA 450/2-78-028a and EPA 450/2-78-0286, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina, May 1978.

4. American Society of Methanical Engineers. "Recommended Guide for the Prediction of Dispersion of Airborne Effluents." M. Smith (editor). American Society of Mechanical Engineers, New York, New York, 1968.

5. Fox, D. G. "Judging Air Quality Model Performance - A Summary of the AMS Workshop on Dispersion Model Performance, Woods Hole, Massachusetts, September 8-11, 1980." Bulletin of the American Meteorological Societh 62:599-609, May 1981.

Preceding page blank

Appendix C

Procedure For Calculating Non-Overlapping Confidence Intervals

This Appendix illustrates the procedure used to calculate non-overlapping confidence intervals as discussed in Section 4.3 of the narrative example. This procedure is used when the 95 percent confidence intervals of the performance measure (absolute value of bias, variance or correlation) contains the value of the performance measure for both models, as illustrated in Figure 2 of Appendix B. The following example demonstrates this procedure.

Suppose that for Model A the value of the bias performance measure is 105 $\mu g/m^3$, the standard error is 20 $\mu g/m^3$, and the sample size is 600; and for Model B, these values are 75 $\mu g/m^3$, 25 $\mu g/m^3$, and 750, respectively. In order to obtain the confidence intervals for the bias of each model, the standard error is multiplied by a factor obtained from a table of the t distribution (This factor is, in turn, a function of the confidence level and the sample size.) The product obtained is then subtracted from the bias to produce the lower confidence bound and added to the bias to produce the upper confidence bound. For this example the 95 percent confidence intervals are 66 $\mu g/m^3$ to 144 $\mu g/m^3$ for Model A and 26 $\mu g/m^3$ to 124 $\mu g/m^3$ for Model B (see Table C-1). In this case the 95 percent confidence interval for each model includes the bias for the other model, but the procedure would be the same if only one of the 95 percent confidence intervals overlapped the bias of the other model.

The procedure next involves relaxing the confidence level until a level is reached at which neither model's confidence interval includes the bias for the other model. Due to the discreet organization of the t-distribution tables, the relaxation was accomplished in stepward decreases from 95 percent confidence to 90 percent, 80 percent, 60 percent, 40 percent, and

Preceding page blank

20 percent, until the non-overlapping level was identified. For the example in Table C-1, the 90 percent confidence intervals are 72 $\mu g/m^3$ to 138 $\mu g/m3$ for Model A and 34 $\mu g/m^3$ to 116 $\mu g/m^3$ for Model B. Again, both confidence

TABLE C-1

EXAMPLE CONFIDENCE INTERVALS AT FOUR CONFIDENCE LEVELS

	Model A Bias = 105 $\mu g/m^3$		Model B Bias = 75 $\mu g/m^3$	
Confidence Level	Lower Bound ($\mu g/m^3$)	Upper Bound ($\mu g/m^3$)	Lower Bound ($\mu g/m^3$)	Upper Bound ($\mu g/m^3$)
95%	66	144	26	124
90%	72	138	34	116
80%	79	131	43	107
60%	88	122	54	96

intervals include the value of the bias for the other model, and therefore the confidence level must be decreased by another step. The 80 percent confidence interval for Model A (79 $\mu g/m^3$ to 131 $\mu g/m^3$) does not include the value of the bias for Model B (75 $\mu g/m^3$). However, since the 80 percent confidence interval for Model B (43 $\mu g/m^3$ to 107 $\mu g/m^3$) does include the value of the bias for Model A (105 $\mu g/m^3$), the confidence level must be decreased by another step. The 60 percent confidence intervals are 88 $\mu g/m^3$ to 122 $\mu g/m^3$ for Model A and 54 $\mu g/m^3$ to 96 $\mu g/m^3$ for Model B. Since neither interval includes the value of the bias for the other model, the non-overlapping confidence interval has been identified as being 60 percent. Thus, in the scoring scheme, 60 percent of the total possible points would be awarded to Model B in the case of this example, the model with the lower bias. For the case when the 20 percent level fails to produce non-overlapping confidence intervals, neither model is awarded any points.